Journal of Biomedical Informatics 44 (2011) 529–535

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An efficient statistical feature selection approach for classification of gene expression data

B. Chandra *, Manish Gupta

Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110 016, India

ARTICLE INFO

Article history:
Received 10 February 2010
Available online 15 January 2011

Keywords:
Cancer diagnosis and prediction
Gene selection
Classification
Feature selection

ABSTRACT

Classification of gene expression data plays a significant role in prediction and diagnosis of diseases. Gene expression data has a special characteristic that there is a mismatch in gene dimension as opposed to sample dimension. All genes do not contribute for efficient classification of samples. A robust feature selection algorithm is required to identify the important genes which help in classifying the samples efficiently. In order to select informative genes (features) based on relevance and redundancy characteristics, many feature selection algorithms have been introduced in the past. Most of the earlier algorithms require computationally expensive search strategy to find an optimal feature subset. Existing feature selection methods are also sensitive to the evaluation measures. The paper introduces a novel and efficient feature selection approach based on statistically defined effective range of features for every class termed as ERGS (*Effective Range based Gene Selection*). The basic principle behind ERGS is that higher weight is given to the feature that discriminates the classes clearly. Experimental results on well-known gene expression datasets illustrate the effectiveness of the proposed approach. Two popular classifiers viz. Nave Bayes Classifier (NBC) and Support Vector Machine (SVM) have been used for classification. The proposed feature selection algorithm can be helpful in ranking the genes and also is capable of identifying the most relevant genes responsible for diseases like leukemia, colon tumor, lung cancer, diffuse large B-cell lymphoma (DLBCL), prostate cancer.

1. Introduction

Terabytes of biological data are being produced at a phenomenal rate using microarray technology. Microarray allows monitoring of thousands of genes in parallel and produce enormous valuable data. Classification is one of the tools in data mining that is being used for classifying samples in gene expression data [2–4,7,20,21,47]. Naive-Bayes Classifier (NBC) [19,9,51], Support Vector Machine (SVM) [48,22], K-Nearest Neighbor (KNN) [1,32] etc. are the commonly used methods of classification for gene expression data.

Earlier studies [5,20,28] depict the importance of feature selection methods for selecting informative genes prior to classification of microarray data for cancer prediction and diagnosis. Feature selection methods [16,23,34] removes irrelevant and redundant features to improve classification accuracy. Feature selection methods can be categorized into filter, wrapper, and embedded or hybrid. The filter approach [29,31] selects features without involving any data-mining algorithm. The filter algorithms are evaluated based on four different evaluation criteria namely, distance, information, dependency and consistency. The wrapper approach [11,30,50] selects feature subset based on the classifier and ranks feature subset using predictive accuracy or cluster goodness. It is more computationally expensive than the filter model. The hybrid model [27] takes advantage of the two models by using their different evaluation criteria in different search stages.

Most of the earlier feature selection methods [11,30,42] can not be used for handling extremely high dimensional gene expression data since they require computationally prohibitive search strategy for finding optimal feature subset. For many real world problems, feature subset generation is an NP-hard problem [30]. To overcome the limitation of existing approaches for gene selection, this paper introduces a novel and efficient feature selection and ranking approach termed as ERGS (*Effective Range based Gene Selection*). ERGS algorithm is based on effective range, which is uniquely defined using statistical inference theory [44,24]. According to statistical inference theory, for a given level of significance, the confidence level is used to indicate the reliability of an interval estimate. The confidence interval of a class distribution may be having wider interval due to the presence of outliers and higher-class variance. The paper introduces statistically defined effective range to overcome the problems of outliers and higher-class variance.

ERGS algorithm has utilized the concept of interval estimate for defining effective ranges of features for every class for a given

* Corresponding author. Present address: Department of Industrial Management and Engineering, Indian Institute of Technology, Kanpur, India.
E-mail address: bchandra104@yahoo.co.in (B. Chandra).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.01.001

feature. The effective range is also based on Chebyshev's inequality, which is true for all distributions. Class prior probability is also taken into account for defining effective range. The feature weights are computed using effective ranges of each class for that feature. A feature is given more weight if the decision boundaries among classes are far apart i.e. the classes can easily be distinguished. It means that the effective ranges of a feature with higher weight do not overlap or have lesser overlapping area. In ERGS algorithm, overlapping area is divided with data range of the feature to scale down the effect of features with higher data range and termed as area coefficient. After computation of area coefficients of all features, these are normalized with maximum area coefficient so that all can be measured on the same scale. The weights for the features are generated by considering the fact that lesser the normalized area coefficient of a feature implies greater is the weight for the feature. A feature is selected if its weight is more than a given threshold value or first few features are selected as significant features after sorting the feature weight in descending order.

The major advantages of ERGS algorithm are that it does not require any search strategy for feature subset generation and iterative process for feature subset evaluation unlike existing feature selection approaches. It can easily be applied for feature selection and ranking in many machine learning problems of classification and clustering. It can also generate feature weights, which can be used in weighted clustering and similar application.

A brief overview of the gene selection methods used in gene expression data analysis is given in Section 2. It also describes two popular gene selection methods considered for comparative evaluation with ERGS algorithm. Details of the proposed ERGS approach for gene selection are given in Section 3. Section 4 presents the brief description of two poplar classification methods used for analysis. Results and discussion are shown in Section 5 that provide the comparative evaluation of proposed approach over popular gene selection algorithms for six well-known gene expression datasets. Concluding remarks are given in the last section of the paper.

2. Existing gene selection methods: a brief overview

In the past, a number of gene selection methods have been introduced to select informative genes for cancer prediction and diagnosis [3,6,18,37–39,41,46]. A comprehensive review of feature selection technique has been described by Saeys et al. [45]. TNoM (threshold number of misclassification) as a score of gene selection is introduced by Ben-Dor et al. [6]. From the TNoM score, a P-value is calculated that represents the significance of each gene. The use of statistical significance t-test and ANOVA [28] has also been used as a criteria for gene selection. Modifications of a number of Bayesian approaches [5,18,53] have also been applied for gene selection in spite of its Gaussian assumptions. Golub et al. [20] have applied clustering for gene expression data [8] to reduce the dimensionality of the data prior to classification. Wong and Hsu [49] has described a two-stage classification method, the first stage being subset-gene-ranking while the second stage deals with classification. Li and Shu [33] proposed a nonlinear dimensionality reduction kernel method and used support vector machine to classify gene expression data. Horng et al. [25] has proposed an expert system to classify gene expression data along with gene selection using decision tree. A sequential feature extraction approach based on stepwise regression and feature transformation using class-conditional independent component analysis was also proposed by Fan et al. [17] for classification of gene expression data.

To highlight the effectiveness of the proposed approach, it has been compared with most commonly used gene selection methods namely, Relief-F [31], minimal-redundancy-maximal relevance (MRMR) [13], *t*-Statistic [36], Information Gain [43,36] and χ^2-Statistic [35,36]. The brief descriptions of algorithms used for comparison are given as follows.

2.1. Relief-F

Relief-F [31] has been introduced as an extension to Relief algorithm [29] for dealing with noisy, incomplete and multi-class datasets. It assigns a "relevance" weight to each feature. Randomly, a sample instance (R) is selected from m sample instances and the relevance values are updated based on the difference between the selected instance (R) and the nearest instances of the same (H) (called nearest hit) and different class ($M(C)$) (called nearest miss of class C). It gives more weight to features that discriminate the instance from neighbors of different classes. The weights are updated by considering average contribution of nearest misses $M(c)$. The average contribution also takes into account of prior probability of each class. The weight of ith feature X_i is updated as follows

$$w_i = w_i - \frac{\Psi(X_i, R, H)}{m} + \sum_{C \neq C_R} \frac{P(C) * \Psi(X_i, R, M(c))}{m} \tag{1}$$

where the function $\Psi(X_i,R,H)$ calculates the distance between sample instance (R) and nearest hit (H) or nearest misses $M(c)$.

2.2. Minimum redundancy-maximum relevance (MRMR)

MRMR proposed by Ding and Peng [13] selects features by minimizing redundancy among them with maximal relevance. MRMR uses mutual information criterion [40,52] as a measures of relevance for discrete datasets whereas F-statistic between the genes and the class variable is considered as the score of maximum relevance for the continuous variables. The experimental results of the current study has been compared with MRMR approaches for continuous variables since gene expression data is of continuous type. The F-test value of feature X_i is defined by

$$F(X_i, C) = \left[\sum_j n_j(\mu_{ij} - \mu_i)/(l-1)\right] \Big/ \sigma^2. \tag{2}$$

where $C = \{C_j\}$ is the class set $j = 1, 2, \ldots, l$, μ_i is the mean of X_i, μ_{ij} denotes the mean of X_i for class C_j, and $\sigma^2 = \left[\sum_j (n_j - 1)\sigma_j^2\right] \Big/ (n - l)$ is the pooled variance for given size n_j and variance σ_j^2 of class C_j.

The maximum relevance criterion, for feature subset, S is given by

$$\max_S \left[\frac{1}{|S|} \sum_{i \in S} F(X_i, C)\right] \tag{3}$$

According to the method, the first feature is selected by Eq. (3) and rest of the features are selected using linear incremental search algorithm based on optimization criterion function. MRMR-FDM(F-test Distance Multiplicative) and MRMR-FSQ (F-test Similarity Quotient) are the two popular linear search schemes for continuous variables. For given feature set $\mathbf{X}$, the optimization condition for MRMR-FDM is defined by

$$\max_{i \in \mathbf{X} - S} \left[F(X_i, C) \cdot \frac{1}{|S|} \sum_{k \in S} d(X_i, X_k)\right] \tag{4}$$

where $d(X_i,X_k)$ is the Euclidean distance between feature X_i and X_k. Similarly, MRMR-FSQ optimization criterion is given by

$$\max_{i \in X - S} \left[F(X_i, C) \Big/ \left[\frac{1}{|S|} \sum_{k \in S} \frac{1}{d(X_i, X_k)}\right]\right] \tag{5}$$

2.3. t-Statistic

This gene selection method utilizes t-Statistic and popular in two class problems i.e. each sample can be classified either into class C_1 or to class C_2. For each feature X_i, t-Statistic is computed by

$$t(X_i) = \frac{|\mu_{i1} - \mu_{i2}|}{\sqrt{\frac{\sigma_{i1}^2}{n_1} + \frac{\sigma_{i2}^2}{n_2}}} \tag{6}$$

where μ_{ij} denotes mean of ith feature X_i for class C_j and σ_{ij} denotes Standard Deviation of ith feature X_i for class C_j. The class index is denoted by j i.e. $j = 1$ or $j = 2$. After calculating the values of t-Statistic for each feature, we sort these values in descending order in order to select the important the feature.

2.4. Information Gain

Information gain is popularly used as attribute selection criteria in Decision Tree by Quinlan [43]. Liu et al. [36] has used it as a gene selection criterion. Let $C = \{C_j\}$ the class set $j = 1, 2, \ldots, l$. For each feature X_i, Information Gain is measured as

$$InfoGain(X_i) = H(C) - H(C/X_i) \tag{7}$$

where

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{8}$$

and

$$H(C/X_i) = -\sum_{x \in X_i} p(x) \sum_{c \in C} p(c/x) \log_2 p(c/x) \tag{9}$$

Information Gain can be used only on discrete features and hence for numeric features discretization is necessary prior to computing Information Gain. Entropy-based discretization method is generally used for gene expression data. Similar, to t-Statistic, features are selected based on the larger values of Information Gain.

2.5. χ^2-Statistic

The value of χ^2-Statistic is computed for each feature individually with respect to the classes. Similar to Information Gain, each numeric attribute is discretized before computing χ^2-Statistic. For each feature X_i, χ^2-Statistic is defined as

$$\chi^2 = \sum_{x \in X_i} \sum_{c \in C} \frac{\left(n_{(x \in X_i \& c \in C)} - e_{(x \in X_i \& c \in C)}\right)^2}{e_{(x \in X_i \& c \in C)}} \tag{10}$$

where $n_{(x \in X_i \& c \in C)}$ is the number of samples in X_i for class c whose value is x. The expected frequency $e_{(x \in X_i \& c \in C)}$ is defined as

$$e_{(x \in X_i \& c \in C)} = \frac{n_{x \in X_i} \times n_{c \in C}}{n} \tag{11}$$

where $n_{x \in X_i}$ denotes the number of samples in X_i with value x and $n_{c \in C}$ represents the number of samples of class c. n is the total number of samples. The features are selected based on the sorted values of χ^2-Statistic for all features.

3. ERGS (*Effective Range based Gene Selection*) algorithm

ERGS algorithm does not require any search strategy for feature subset generation unlike most of the popular feature selection algorithm. It does not require iterative process for feature subset evaluation criterion as the case of many existing feature selection algorithm. It works under the principle that a feature should be given more weight if decision boundaries among classes are very far away from each other i.e. classes can easily be distinguished. The decision boundaries of the classes are obtained by statistically defined effective range.

3.1. Effective range (R_{ij})

Definition 1. Let $X = \{X_1, X_2, \ldots, X_d\}$ be the feature set of dataset D of size n and dimension d. $C = \{C_j\}$ be the class set $j = 1, 2, \ldots, l$. p_j is the class probability of jth class C_j. For each class C_j of ith feature X_i, μ_{ij} denotes mean of ith feature X_i for class C_j and σ_{ij} denotes Standard Deviation of ith feature X_i for class C_j. Effective Range (R_{ij}) of jth class C_j for ith feature X_i is defined by

$$R_{ij} = [r_{ij}^-, r_{ij}^+] = [\mu_{ij} - (1 - p_j)\gamma\sigma_{ij},\ \mu_{ij} + (1 - p_j)\gamma\sigma_{ij}] \tag{12}$$

where r_{ij}^- and r_{ij}^+ are the lower and upper bounds of the effective range respectively. p_j is the prior probability of jth class C_j. Here, the factor $(1 - p_j)$ is taken to scale down effect of class with high probabilities and consequently large variance. The details of effect of the factor $(1 - p_j)$ are given in Section 3.3. The value of γ is determined statistically by Chebyshev Inequality defined as

$$P(|X - \mu_{ij}| \geqslant \gamma\sigma_{ij}) \leqslant \frac{1}{\gamma^2} \tag{13}$$

Which is true for all distributions.

The value of γ is computed as 1.732 for the effective range which contains at least 2/3rd of the data objects.

3.2. ERGS algorithm

The steps for ERGS algorithm is described as follows

1. Calculate Effective Ranges (R_{ij}) for all classes of each feature X_i using Eq. (12)
2. Sort the effective ranges of classes in ascending order to compute Overlapping Area (OA_i) for each feature X_i.
3. Compute Overlapping Area (OA_i) among classes of feature X_i as

$$OA_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} \varphi_i(j,k) \tag{14}$$

where $\varphi_i(j,k) = \begin{cases} r_{ij}^+ - r_{ik}^- & \text{if } r_{ij}^+ > r_{ik}^- \\ 0 & \text{otherwise} \end{cases}$

4. Compute Area Coefficient (AC_i) of feature X_i as

$$AC_i = \frac{OA_i}{Max_j(r_{ij}^+) - Min_j(r_{ij}^-)} \tag{15}$$

5. Compute Normalized Area Coefficient (NAC_i)

$$NAC_i = AC_i / Max(AC_j), for j = 1, \ldots, d \tag{16}$$

6. Compute Weight (w_i) of ith feature X_i as

$$w_i = 1 - NAC_i \tag{17}$$

7. Select feature X_i, if $w_i > \theta$, where θ is threshold value.

It is to be noticed that the different features are distributed on different data range i.e. some may be from 0 to 10 and another may be from 0-10000. It means that OA may be larger for those features, which are having higher data range. To nullify this effect, (OA) is divided with data range of the feature. The subsequent subsections deal with some theoretical aspects of ERGS algorithm.

Theorem 1. *Let $f_{ij}(x)$ be the probability density function of the class distribution of the jth class C_j for ith feature X_i denoted by $\Gamma_{ij}(\mu_{ij}, \sigma_{ij})$, where μ_{ij} and σ_{ij} are mean and standard deviation of ith feature X_i for class C_j respectively. Then*

$$P(R_{ij}) \geqslant 1 - \frac{1}{(1-p_j)^2\gamma^2} \tag{18}$$

Proof. By definition, for any class distribution Γ_{ij}

$$\sigma_{ij}^2 = \int_{-\infty}^{\infty}(x-\mu_{ij})^2 f_{ij}(x)\,dx \geqslant \int_{\mu_{ij}-(1-p_j)\gamma\sigma_{ij}}^{\mu_{ij}+(1-p_j)\gamma\sigma_{ij}}(x-\mu_{ij})^2 f_{ij}(x)\,dx$$
$$\geqslant \gamma^2(1-p_j)^2\sigma_{ij}^2\int_{\mu_{ij}-(1-p_j)\gamma\sigma_{ij}}^{\mu_{ij}+(1-p_j)\gamma\sigma_{ij}} f_{ij}(x)\,dx \geqslant \gamma^2(1-p_j)^2\sigma_{ij}^2 P(R_{ij})$$

where, $R_{ij} = [\mu_{ij} - (1-p_j)\gamma\sigma_{ij} \leqslant x \leqslant \mu_{ij} + (1-p_j)\gamma\sigma_{ij}]$

$$P(R_{ij}) \leqslant \frac{1}{\gamma^2(1-p_j)^2}$$

$$P(R_{ij}) \geqslant 1 - \frac{1}{(1-p_j)^2}\gamma^2$$

This inequality is true for any class distribution Γ_{ij}. □

3.3. Effect of factor $(1 - p_j)$

By theorem (1), $P(R_{ij}) \geqslant 1 - \frac{1}{(1-p_j)^2\gamma^2}$ i.e.

$$P(\mu_{ij} - (1-p_j)\gamma\sigma_{ij} \leqslant x \leqslant \mu_{ij} + (1-p_j)\gamma\sigma_{ij}) \geqslant 1 - \frac{1}{\gamma^2}(1-p_j)^2 \tag{19}$$

If σ_{ij} is increased then R_{ij} is also increased and consequently Overlapping Area(OA_i) among classes of Feature X_i [defined by Eq. (14)] and Area Coefficient (AC_i) of Feature X_i [defined by Eq. (15)] are increased. That leads to decrease in weight (w_i) of Feature X_i. Therefore, R_{ij} is measured such that it maximizes weight (w_i) of Feature X_i. In the proposed approach, σ_{ij} is scaled down by $(1-p_j)$, then, since $0 < p_j < 1$, It means that Effective Range (R_{ij}) is reduced using, which not only gives more weight (w_i) of Feature X_i but also nullify the effect of outliers for computing weight. The effect of proposed effective range is further analyzed using classical statistical inference theory like Discriminant Rule and Maximum Likelihood Rule. It is found from the experimental results that Expected Cost of Misclassification (*ECM*) is decreased using ERGS algorithm.

3.4. An example

Illustration of ERGS algorithm has been done using MLL(Mixed-Lineage Leukaemia)[4], a benchmark gene dataset. MLL dataset contains 72 samples of 12,562 genes. The samples consist of three types of leukaemia namely, acute lymphoblastic leukaemia (ALL), Mixed-Lineage Leukaemia (MLL) and acute myeloblastic leukaemia (AML).

In ERGS algorithm, more weight is assigned to a gene if it discriminates the classes clearly i.e. it has less overlapping area. Using ERGS algorithm, the weight for gene with Gene Accession No. '32864_at' is computed as 0.90 which shows high weightage for this gene. This justifies the fact that there is no overlapping area between AML and ALL types of leukaemia for gene '32864_at' as shown in Fig. 1. It means that gene '32864_at' can not have any ambiguity to classify AML and ALL types of leukaemia. Therefore, it is shown in the example that ERGS algorithm gives more weight to those features that are helpful for classifying the data accurately

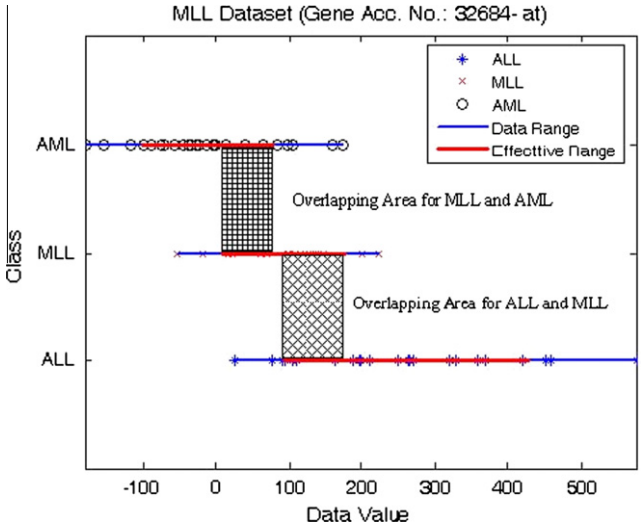

Fig. 1. Data Plot of MLL dataset: Gene Accession No. 32684_at.

i.e. the feature should not lead to any ambiguity in the classification process. The features with higher weights also clearly describe the decision boundary in the classification process.

4. Classification methods used for analysis

4.1. Naive-Bayes Classifier (NBC)

NBC [19] is a simple probabilistic classifier with assumption of independence among attributes. Domingos and Pazzani [14] has also found that this assumption has less impact than might be expected. It often provides better classification accuracy on gene expression data than any other classifier does. NBC learns from training data and then predicting the class of the test instance with the highest posterior probability. Let C be the random variable that denotes the class of an instance and let $X\langle X_1, X_2, \ldots, X_m\rangle$ be a vector of random variables denoting the observed attribute values. Let c_j represent jth class label and let $x\langle x_1, x_2, \ldots, x_m\rangle$ represent a particular observed attribute value vector. To predict the class of a test instance **x**, Bayes' theorem is used to compute the probability as follows:

$$p(C = c_j | X = \mathbf{x}) \propto p(C = c_j)\prod_{i=1}^{m} p(X_i = x_i | C = c_j) \tag{20}$$

Then, the class of test instance is predicted to the class with highest probability. Here $X = \mathbf{x}$ represents the event that $X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots X_m = x_m$. For test instances, it can easily be computed using training data. In this paper, Equal width discretization (EWD) [15,51] is used to transform numeric attributes into discrete one in NBC. A discrete attribute X_i^c is formed for each numeric attribute X_i and each value of X_i^c corresponds to an interval $(a_i, b_i]$ of X_i. If $x_i(a_i, b_i]$, then in (3) is estimated by

$$p(C = c_j | X = x) \propto p(C = c_j)\prod_{i=1}^{m} p(a_i < x_i \leqslant b_i | C = c_j) \tag{21}$$

However, for high dimensional gene expression data underflow limitation occurs because the multiplication of large number of probabilities can result in floating-point underflow [9]. Logarithm function is used to address this problem of underflow. All computations are performed by summing logs of probabilities rather than multiplying the probabilities. Class with highest final log probability score is still the most probable.

4.2. Support Vector Machine (SVM)

SVM [48] performs classification by constructing optimal hyperplanes in the feature vector space to maximize the margin between a set of objects of different classes. To construct an optimal hyperplane, an iterative training algorithm is used to minimize an error function $\Lambda(w)$ defined by

$$\Lambda(w) = \frac{1}{2} w^T w + C \sum \xi_i \tag{22}$$

subject to the constraints:

$$y_i[w^T \mathbf{K}(x_i) + b] \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0, \quad i = 1, \ldots, n \tag{23}$$

where w is the vector of coefficients, b a constant and ξ_i, $i = 1, \ldots, n$ are the parameters to allow misclassification of difficult or noisy examples. For each training instance i, x_i are the independent variables represented by class labels y_i. The kernel function $\mathbf{K}$ transforms input data into higher-dimensional feature space. It is used to construct nonlinear decision boundary. For the current study, linear kernel function is used for transformation. The parameter C can be viewed as a way to control overfitting. The larger the value of C, the more the error is penalized.

Standard SVM can be applied for 2 class problems. The multi-class problems are solved either by constructing a multi-class classifier using binary classifiers such as one-against-others or all-against-all [12] or by applying directly a multi-class SVM [26]. In this paper, the results of SVM classification have been obtained using publicly available Matlab version of LIBSVM [10].

5. Results and discussion

The performance of the proposed feature selection algorithm, ERGS (*Effective Range based Gene Selection*) has been evaluated on six well-known gene expression datasets namely ALL_AML [20], colon tumor [3], diffuse large B-cell lymphoma (DLBCL) [2], lung cancer [21], mixed-lineage leukaemia (MLL) [4] and prostate [47]. Classification of the samples in the gene expression datasets on the selected features(using ERGS algorithm) was done using Nave Bayes Classifier (NBC) and Support Vector Machine (SVM). Details of the gene datasets are as follows.

Table 1
LOOCV classification accuracies with NBC of six gene expression datasets for different gene selection methods using 10–100 selected genes.

Dataset	Method	NBC					
		10	20	40	60	80	100
ALL_AML	ERGS	98.61	97.22	97.22	97.22	97.22	97.22
	Relief-F	93.06	91.67	94.44	91.67	91.67	93.06
	MRMR-FDM	58.33	68.06	61.11	70.83	65.28	65.28
	MRMR-FSQ	48.61	65.28	62.50	58.33	66.67	65.28
	t-Statistic	94.44	95.83	97.22	97.22	97.22	97.22
	Info. Gain	94.44	97.22	95.83	95.83	95.83	95.83
	χ^2-Statistic	97.22	97.22	95.83	95.83	95.83	95.83
COLON	ERGS	82.26	82.26	79.03	80.65	79.03	83.87
	Relief-F	70.97	75.81	75.81	74.19	75.81	79.03
	MRMR-FDM	46.77	46.77	53.23	56.45	61.29	66.13
	MRMR-FSQ	51.61	48.39	58.06	59.68	64.52	64.52
	t-Statistic	82.26	77.42	79.03	80.65	79.03	79.03
	Info. Gain	79.03	79.03	77.42	80.65	79.03	82.26
	χ^2-Statistic	80.65	79.03	79.03	77.42	79.03	79.03
DLBCL	ERGS	94.79	92.71	94.79	94.79	93.75	93.75
	Relief-F	93.75	90.63	90.63	92.71	91.67	90.63
	MRMR-FDM	90.63	89.58	88.54	90.63	91.67	91.67
	MRMR-FSQ	82.29	90.63	90.63	90.63	90.63	91.67
	t-Statistic	93.75	91.67	93.75	94.79	93.75	93.75
	Info. Gain	92.71	92.71	92.71	92.71	92.71	92.71
	χ^2-Statistic	94.79	91.67	93.75	93.75	93.75	93.75
LUNG	ERGS	95.03	96.13	98.90	98.90	98.34	100.00
	Relief-F	92.82	95.03	92.27	97.79	97.24	98.34
	MRMR-FDM	83.43	88.40	91.71	92.82	92.27	92.82
	MRMR-FSQ	82.87	83.43	90.06	90.06	90.06	91.71
	t-Statistic	92.82	92.82	97.24	97.24	97.79	97.79
	Info. Gain	93.37	93.37	93.37	95.03	95.03	95.03
	χ^2-Statistic	92.82	93.37	93.37	95.03	95.03	95.03
MLL	ERGS	94.44	94.44	94.44	95.83	95.83	97.22
	Relief-F	93.06	90.28	90.28	88.89	88.89	90.28
	MRMR-FDM	40.28	41.67	47.22	50.00	47.22	50.00
	MRMR-FSQ	43.06	34.72	54.17	50.00	50.00	48.61
	Info. Gain	93.06	94.44	95.83	94.44	95.83	94.44
	χ^2-Statistic	90.28	93.06	94.44	95.83	94.44	94.44
PROSTATE	ERGS	94.12	93.14	92.16	92.16	92.16	91.18
	Relief-F	64.71	82.35	81.37	79.41	81.37	80.39
	MRMR-FDM	56.86	53.92	61.77	62.75	64.71	63.73
	MRMR-FSQ	47.06	61.77	61.77	63.73	63.73	63.73
	t-Statistic	93.14	91.18	92.16	92.16	91.18	91.18
	Info. Gain	94.12	93.14	91.18	91.18	91.18	91.18
	χ^2-Statistic	91.18	92.16	91.18	91.18	91.18	91.18

Table 2
LOOCV classification accuracies with SVM of six gene expression datasets for different gene selection methods using 10 to 100 selected genes.

Dataset	Method	SVM					
		10	20	40	60	80	100
ALL_AML	ERGS	93.06	97.22	97.22	98.61	100.00	98.61
	Relief-F	81.94	90.28	84.72	86.11	87.50	93.06
	MRMR-FDM	58.33	61.11	70.83	80.56	84.72	81.94
	MRMR-FSQ	48.61	59.72	77.78	84.72	87.50	80.56
	t-Statistic	91.67	97.22	95.83	98.61	98.61	97.22
	Info. Gain	91.67	94.44	95.83	98.61	98.61	97.22
	χ^2-Statistic	91.67	95.83	95.83	98.61	97.22	97.22
COLON	ERGS	82.26	80.65	79.03	82.26	80.65	83.87
	Relief-F	69.35	75.81	66.13	75.81	77.42	75.81
	MRMR-FDM	66.13	70.97	70.97	66.13	62.90	67.74
	MRMR-FSQ	62.90	70.97	66.13	69.35	67.74	67.74
	t-Statistic	79.03	77.42	74.19	72.58	74.19	80.65
	Info. Gain	77.42	79.03	75.81	79.03	77.42	77.42
	χ^2-Statistic	79.03	79.03	77.42	74.19	75.81	79.03
DLBCL	ERGS	92.71	93.75	95.83	96.88	96.88	95.83
	Relief-F	91.67	89.58	89.58	85.42	92.71	92.71
	MRMR-FDM	91.67	90.63	91.67	94.79	94.79	93.75
	MRMR-FSQ	82.29	89.58	90.63	89.58	93.75	94.79
	t-Statistic	96.88	95.83	95.83	95.83	96.88	95.83
	Info. Gain	96.88	96.88	96.88	96.88	96.88	97.92
	χ^2-Statistic	96.88	95.83	97.92	96.88	97.92	95.83
LUNG	ERGS	98.34	98.34	99.45	99.45	99.45	99.45
	Relief-F	97.24	97.24	98.90	98.90	98.90	98.90
	MRMR-FDM	82.87	86.74	87.29	91.16	95.58	95.58
	MRMR-FSQ	83.43	87.29	82.87	87.29	91.71	93.37
	t-Statistic	97.79	97.24	97.79	98.34	99.45	99.45
	Info. Gain	98.34	97.24	99.45	99.45	98.90	98.90
	χ^2-Statistic	98.34	95.03	99.45	99.45	98.90	99.45
MLL	ERGS	88.89	93.06	97.22	95.83	95.83	97.22
	Relief-F	80.56	87.50	87.50	88.89	93.06	91.67
	MRMR-FDM	59.72	54.17	59.72	72.22	73.61	79.17
	MRMR-FSQ	44.44	56.94	65.28	69.44	69.44	66.67
	Info. Gain	87.50	91.67	91.67	95.83	97.22	97.22
	χ^2-Statistic	87.50	93.06	93.06	90.28	91.67	95.83
PROSTATE	ERGS	89.22	89.22	91.18	93.14	93.14	93.14
	Relief-F	87.25	84.31	86.27	88.24	88.24	93.14
	MRMR-FDM	67.65	62.75	61.76	60.78	71.57	72.55
	MRMR-FSQ	60.78	68.63	63.73	62.75	73.53	73.53
	t-Statistic	90.20	84.31	87.25	91.18	93.14	92.16
	Info. Gain	87.25	87.25	86.27	88.24	89.22	88.24
	χ^2-Statistic	89.22	89.22	90.20	91.18	90.20	86.27

Table 3
The top 10 selected genes using ERGS algorithm for Colon data.

Gene accession number	Gene description
'H08393'	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
'X63629'	H.sapiens mRNA for p cadherin.
'M22382'	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
'R36977'	P03001 TRANSCRIPTION FACTOR IIIA.
'T56604'	TUBULIN BETA CHAIN (Haliotis discus)
'H40095'	MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
'U30825'	Human splicing factor SRp30c mRNA, complete cds
'T47377'	S-100P PROTEIN (HUMAN)
'J05032'	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
'M63391'	Human desmin gene, complete cds

5.1. Datasets description

5.1.1. ALL_AML

ALL_AML data consists of 72 samples. The samples consist of two types of leukaemia, 25 samples of acute myeloblastic leukaemia (AML) and 47 samples of acute lymphoblastic leukaemia (ALL). The samples are taken from 63 bone marrow samples and 9 peripheral blood samples. There are 7192 genes in the dataset.

5.1.2. Colon tumor

Colon dataset consists of 62 samples of colon epithelial cells from colon-cancer patients. The samples consist of tumor bipsies collected from tumors, and normal biopsies collected from healthy part of the colons of the same patient. The number of genes in the dataset is 2000.

5.1.3. Diffuse large B-cell lymphoma (DLBCL)

DLBCL is an aggressive malignancy of mature B- lymphocytes. DLBCL dataset contains 96 samples of 4026 genes. The samples are categorized into two molecularly distinct forms of DLBCL i.e. 'germinal centre B-like DLBCL' and 'activated B-like DLBCL'.

5.1.4. Lung cancer

Lung dataset contains 181 tissue samples described by 12,533 genes and categorized into 2 classes namely, malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). The 181 samples include 31 MPM and 150 ADCA.

5.1.5. Mixed-Lineage Leukaemia (MLL)

MLL dataset contains 72 samples of 12,582 genes. The samples consist of three types of leukaemia, 24 samples of acute lymphoblastic leukaemia (ALL), 20 samples of Mixed-Lineage Leukaemia (MLL), and 28 samples of acute myeloblastic leukaemia (AML).

5.1.6. Prostate

Prostate cancer dataset contains 102 samples of 12,600 genes and categorized into two classes. The dataset contains gene expression patterns from 52 tumor and 50 normal prostate samples.

5.2. Comparative evaluation

In order to find the efficiency of the ERGS algorithm, the classification accuracy obtained using this algorithm is compared with the accuracy of the other feature selection algorithm. *Leave one out cross validation* (LOOCV) has been used as the validation strategy to give a relatively comprehensive comparison on the performances. The comparative evaluation of ERGS, Relief-F, two schemes of MRMR for continuous variables namely, MRMR-FDM and MRMR-FSQ algorithm, *t*-Statistic, Information Gain and χ^2-Statistic have been carried out for different subsets of selected genes from 10 to 100 for all the datasets. Since *t*-Statistic can be applied only two class problem therefore it can not be used for MLL dataset. Table 1 presents classification accuracies obtained for the above mentioned six gene expression datasets for different gene selection approaches. The six gene subsets of top 10, 20, 40, 60, 80 and 100 genes are selected to highlight the effectiveness of ERGS over other gene selection methods. The reason for superior performance of ERGS selected genes as opposed of others for every gene subsets is that gene selection by ERGS is more efficient and robust. Even by selecting merely top 10 genes, ERGS algorithm is able to achieve 98.61% classification accuracy using NBC for ALL_AML dataset. It is shown in Table 1 that ERGS algorithm is able to select the best informative genes for classification as compared to other feature selection techniques.

For ALL_AML gene expression data, there is a substantial improvement in classification accuracy using ERGS algorithm for every feature subsets starting from 10 to 100. It demonstrates the fact that ERGS is able to select the best informative genes as compared to other well-known techniques. For very high dimensional gene expression datasets like MLL, lung cancer etc. classifier with ERGS selected gene subsets can classify with high degree of accuracy. ERGS not only improves the classification accuracy for gene expression data but also identifies the informative genes responsible for diseases like leukemia, colon tumor, lung cancer, DLBCL and prostate cancer.

It is also seen from Tables 1 and 2 that the classification accuracy using ERGS for feature selection has significantly better than popular feature selection algorithm like *t*-Statistic, Information Gain and χ^2-Statistic. The classification accuracy using ERGS is also consistently improved for different subsets of selected genes. ERGS algorithm performs remarkably well as compared to other feature selection algorithms for prostate dataset.

Tables 3 and 4 present gene accession number and gene description of the top ten selected genes by ERGS algorithm for colon data and MLL data respectively. The results shown in Tables 3 and 4 are commensurate with the clinically proven results.

Table 4
The top 10 selected genes using ERGS algorithm for MLL data.

Gene accession number	Gene description
'32847_ at'	Hs.211582 gnl|UG|Hs#S417769 Homo sapiens myosin light chain kinase (MLCK) mRNA, complete cds
'1389_ at'	Hs.1298 gnlUGHs#S1945 Human common ALL antigen (CALLA) mRNA, complete cds
'36239_ at'	Hs.2407 gnlUGHs#S226199 H.sapiens mRNA for oct-binding factor
'37539_ at'	Hs.79219 gnlUGHs#S1569334 Homo sapiens mRNA for KIAA0959 protein, partial cds
'39931_ at'	Hs.38018 gnlUGHs#S952957 Homo sapiens mRNA for protein kinase, Dyrk3
'266_ s_ at'	Hs.278667 gnlUGHs#S885 Homo sapiens CD24 signal transducer mRNA, complete cds and 3'' region
'32579_ at'	Hs.78202 gnlUGHs#S6032 Human transcriptional activator (BRG1) mRNA, complete cds
'963_ at'	Hs.166091 gnlUGHs#S5776 H.sapiens mRNA for DNA ligase IV
'35260_ at'	Hs.52081 gnlUGHs#S1367627 Homo sapiens mRNA for KIAA0867 protein, complete cds
'32872_ at'	Hs.202685 gnlUGHs#S1368108 Homo sapiens mRNA; cDNA DKFZp564I083 (from clone DKFZp564I083)

6. Conclusions

In this paper, ERGS (*Effective Range based Gene Selection*), a novel statistical approach of feature selection and ranking has been proposed in order to select informative genes for classifying gene expression data. The governing principle for ERGS algorithm is based on the fact that a feature should be given higher weightage if it discriminates the classes clearly. ERGS algorithm is based on statistically defined effective ranges of each class for a given feature. ERGS algorithm also nullifies the effect of outliers and classes with large variance. ERGS algorithm does not require any computationally extensive search strategy and evaluation criteria unlike many feature selection algorithms. ERGS algorithm is fast, easy to implement and does not require any distribution assumption. The effectiveness of ERGS algorithm has been illustrated using six well-known gene expression datasets. The results confirm that ERGS is a promising feature selection algorithm for gene expression data analysis. ERGS algorithm performs remarkably well in terms of classification accuracy for gene expression datasets as compared to existing popular feature selection algorithm like Relief-F, minimal-redundancy-maximal relevance (MRMR), *t*-Statistics, Information Gain and χ^2-Statistic. ERGS algorithm can be applied for finding weights for features in most of the practical problems, where features do not have equal weights. The approach can play an important role for wider variety of pattern recognition and machine learning problems. It can also be used for very high dimensional datasets like spam filter and document classification.

Acknowledgments

We are thankful to anonymous reviewers for their valuable comments. The comments are helpful for improving the manuscript.

References

[1] Aha D, Kibler D. Instance-based learning algorithms. Mach Learn 1991;6:37–66.
[2] Alizadeh AA et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
[3] Alon U et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745–50.
[4] Armstrong SA et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41–7.
[5] Baldi P, Long A. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001;17:509–16.
[6] Ben-Dor A et al. Tissue classification with gene expression profiles. J Comput Biol 2000;7:559–84.
[7] Bhattacharjee A et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98:13790–5.
[8] Chandra B et al. A new approach: interrelated two way clustering of gene expression data. Stat Methodol 2006;3:93–102.
[9] Chandra B. Robust approach for estimating probabilities in Naive-Bayes classifier. Lecture notes in computer science (LNCS), vol. 4815; 2007. p. 11–6.
[10] Chang CC, Lin CJ. LIBSVM: a library for support vector machines; 2001. <http://www.csie.ntu.edu.tw/cjlin/libsvm>.
[11] Devijver P, Kittler J. Pattern recognition: a statistical approach. Prentice Hall; 1982.
[12] Ding C, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349–58.
[13] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3:185–205.
[14] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 1997;29:103–30.
[15] Dougherty J et al. Supervised and unsupervised discretization of continuous features. In: Proc. 12th int. conf. on machine learning; 1995. p. 194–202.
[16] Dy JG, Brodley CE. Feature subset selection and order identification for unsupervised learning. In: Proc. 17th int. conf. on machine learning; 2005. p. 247–54.
[17] Fan L et al. A sequential feature extraction approach for nave Bayes classification. Expert Syst Appl 2009;36(6):9919–23.
[18] Fox R, Dimmic M. A two-sample Bayesian t-test for microarray data. BMC Bioinform 2006;7:126.
[19] Friedman N et al. Bayesian network classifiers. Mach Learn 1997;29:131–63.
[20] Golub TR et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[21] Gordon GJ et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7.
[22] Guyon I et al. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
[23] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[24] Härdle W, Simar L. Applied multivariate statistical analysis. Springer-Verlag; 2007.
[25] Horng JT et al. An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst Appl 2009;36(5):9072–81.
[26] Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines. IEEE Trans Neural Netw 2002;13:415–25.
[27] Huang J et al. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognit Lett 2007;28:1825–44.
[28] Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak 2006;6:27.
[29] Kira K, Rendell LA. A practical approach to feature selection. In: Proc. 9th int. conf. on machine learning; 1992. p. 249–56.
[30] Kohavi R, John G. Wrapper for feature selection. Artif Intell 1997;97:273–324.
[31] Kononenko I. Estimating features: analysis and extension of RELIEF. In: Proc. 6th European conf. on machine learning; 1994. p. 171–82.
[32] Li L et al. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17:1131–42.
[33] Li X, Shu L. Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. Expert Syst Appl 2009;36(4):7644–50.
[34] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Boston: Kluwer Academic; 1998.
[35] Liu H, Setiono R. Chi2: feature selection and discretization of numeric attributes. In: Proc. 7th international conference on tools with artificial intelligence, Herndon, VA; 1995. p. 388–91.
[36] Liu H et al. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform 2002;13:51–60.
[37] Liu H et al. Ensemble gene selection by grouping for microarray data classification. J Biomed Inform 2010;43:81–7.
[38] Lu Y, Han J. Cancer classification using gene expression data. Inform Syst 2003;28(4):243–68.
[39] Ooi C, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 2003;19:37–44.
[40] Peng H et al. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38.
[41] Peng Y et al. A novel feature selection approach for biomedical data classification. J Biomed Inform 2010;43:15–23.
[42] Pudil P et al. Floating search methods in feature selection. Pattern Recognit Lett 1994;15:1119–25.
[43] Quinlan JR. Induction of decision trees. Mach Learn 1986:81–106.
[44] Rao CR. Linear statistical inference and its application. John Wiley and Sons; 1965.
[45] Saeys Y et al. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:25072517.
[46] Shen Q et al. New gene selection method for multiclass tumor classification by class centroid. J Biomed Inform 2009;42:59–65.
[47] Singh D et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–9.
[48] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
[49] Wong TT, Hsu CH. Two-stage classification methods for microarray data. Expert Syst Appl 2008;34:375–83.
[50] Xiong M et al. Biomarker identification by feature wrappers. Genome Res 2001;11:1878–87.
[51] Yang Y, Webbu G. A comparative study of discretization methods for Naive-Bayes classifiers. In: Proc. pacific rim knowledge acquisition workshop, Japan; 2002. p. 159–73.
[52] Zhang Y et al. Gene selection algorithm by combining reliefF and mRMR. BMC Genom 2007;9(Suppl 2):S27.
[53] Zhoua X et al. Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform 2004;37:249–59.