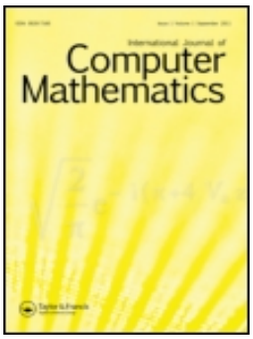

International Journal of Computer Mathematics

ISSN: 0020-7160 (Print) 1029-0265 (Online) Journal homepage: https://www.tandfonline.com/loi/gcom20

An incremental feature selection approach based on scatter matrices for classification of cancer microarray data

Manju Sardana, R.K. Agrawal & Baljeet Kaur

To cite this article: Manju Sardana, R.K. Agrawal & Baljeet Kaur (2015) An incremental feature selection approach based on scatter matrices for classification of cancer microarray data, International Journal of Computer Mathematics, 92:2, 277-295, DOI: 10.1080/00207160.2014.905680

To link to this article: https://doi.org/10.1080/00207160.2014.905680

Published online: 21 May 2014.

Submit your article to this journal

Article views: 181

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gcom20

International Journal of Computer Mathematics, 2015
Vol. 92, No. 2, 277–295, http://dx.doi.org/10.1080/00207160.2014.905680

Taylor & Francis
Taylor & Francis Group

An incremental feature selection approach based on scatter matrices for classification of cancer microarray data

Manju Sardana[a*], R.K. Agrawal[a] and Baljeet Kaur[b]

[a]*School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India;*
[b]*Hansraj College, University of Delhi, Delhi 110007, India*

(*Received 22 September 2012; revised version received 11 March 2014; accepted 12 March 2014*)

Microarray data are often characterized by high dimension and small sample size. There is a need to reduce its dimension for better classification performance and computational efficiency of the learning model. The minimum redundancy and maximum relevance (mRMR), which is widely explored to reduce the dimension of the data, requires discretization and setting of external parameters. We propose an incremental formulation of the trace of ratio of the scatter matrices to determine a relevant set of genes which does not involve discretization and external parameter setting. It is analytically shown that the proposed incremental formulation is computationally efficient in comparison to its batch formulation. Extensive experiments on 14 well-known available microarray cancer datasets demonstrate that the performance of the proposed method is better in comparison to the well-known mRMR method. Statistical tests also show that the proposed method is significantly better when compared to the mRMR method.

Keywords: microarrays; feature selection; minimum redundancy maximum relevance; ratio of scatter matrices; cancer classification

2010 AMS Subject Classifications: 03D15; 15A09; 68T10; 62H30

1. Introduction

Microarray technology allows simultaneous measurement of thousands of genes. This has revolutionized the research that involves the molecular basis of disease diagnosis. Many data mining and machine-learning techniques [20,29,36,52] have been extensively applied for classification of cancer microarray data. Cancer microarray classification refers to the prediction of the cancer class i.e. assignment of a given sample to a predefined cancer class where a class represents a particular cancer type such as acute lymphoblastic leukaemia or acute myeloid leukaemia. In general, the microarray data are characterized by a large number of genes and a small number of available samples and hence suffers from the limitation called the 'curse of dimensionality' [4]. Because of the less number of available samples, classification of such data may suffer from the problem of over-fitting. This problem can be overcome by identifying a smaller number of genes accountable for a given disease. Researchers [26] have also shown that the presence of redundant/irrelevant genes may deteriorate the performance of a classifier significantly. The redundant/irrelevant genes increase data acquisition costs and learning time. Hence dimensionality reduction is a crucial step towards the removal of irrelevant and redundant genes and in the identification of a set of relevant

*Corresponding author. Email: manjusardana12@yahoo.co.in

© 2014 Taylor & Francis

genes responsible for a particular disease. This also helps in improving cancer diagnosis and in catalysing drug discovery.

Dimensionality reduction can be done in two ways [21]: feature extraction and feature selection. Feature extraction methods like the principal component analysis (PCA) utilize all the information present in the measurement space to obtain a new transformed space thereby reducing redundancy and selecting important features from the transformed space. Recently, a PCA-based feature extraction framework for dimensionality reduction has been suggested by Pang *et al*. [42] for human detection in image/video. Although the transformed features provide better representation of the data, the features in the new transformed space are unable to preserve the physical interpretation of the underlying data. Feature selection refers to reducing the dimensionality of the feature space by eliminating noisy, irrelevant and redundant features. It results in reduced measurement costs and the selected feature subset is more interpretable since the features/genes are from the original set and therefore are more meaningful in understanding the physical process related to these features. For disease diagnosis, one is not only interested in classifying the samples based on gene expressions but also in identifying the discriminatory features/genes responsible for a particular disease. Hence dimensionality reduction in cancer microarray classification domain is normally carried out with feature/gene selection rather than with feature extraction.

The feature selection algorithm convolves with the classifier to help learn the decision model wherein learning refers to the formulation of a model to reduce the classification error on training data [14]. The involved classifier optimizes its specific criteria function to minimize (maximize) the classification error (accuracy). This may introduce a learning bias. The learning bias can be described as 'any basis for choosing one generalization over another, other than strict consistency with the observed training instances' [39].

A large number of approaches for feature selection have been proposed in literature. These broadly fall into two categories [21,31]: filter approach and wrapper approach. The filter methods employ statistical characteristics of the data for feature selection. The selection is classifier independent and hence less computation-intensive. The filter approach does not take into account the learning bias (effects of the selected feature subset on the classification accuracy of the learning algorithm) introduced by the learning algorithm, so it may not be able to select the most relevant set of features for the learning algorithm. The wrapper methods on the other hand are computationally more expensive since a classifier must be trained for each candidate subset to find features suited to the predetermined learning algorithm. In literature researchers have observed that many times the computationally intensive wrapper methods do not outperform simple filter methods [23]. Moreover, the wrapper approach is computationally feasible only for middle- or low-dimensional data.

Because of their simplicity and less computational burden, filter feature selection approaches have been widely investigated [3,17,31,36,43,48]. In literature, two categories are defined for filter feature selection: univariate and multivariate. The univariate feature selection methods evaluate the relevance of each feature individually. Most of the univariate methods rank features based on the information content or some quality index. They are simple and fast therefore most widely used [5,20,29,56]. Recently, rank canonical correlation analysis [27] has been suggested as an effective and practical ranking algorithm for dimensionality reduction. However, these ranking algorithms assume that the features are independent of each other and hence these algorithms do not take care of redundancy among selected features. One of the popular univariate methods is based on the computation of a score that measures the mutual information between each gene and the corresponding class label. The $m-$top genes with higher scores are considered more discriminatory and thus form a set of relevant genes. However, it has been observed in literature that genes that perform better individually, do not necessarily perform better in combination, as, such set of genes may be correlated.

A good gene selection method not only selects the relevant genes but also reduces the redundancy among the selected gene subset. In literature, some multivariate methods have been suggested to reduce redundancy among the selected set of genes [3,33,43]. The multivariate approach evaluates dependency between a set of features and the corresponding class variable [6,8,22]. One of the widely employed and popular methods is the minimum redundancy maximal relevance (mRMR). It provides a minimal subset of the non-redundant and relevant genes. However, the method requires discretization of the features before the feature selection procedure, but at the time of evaluation, the original continuous values are used. Since the choice of the discretization strategy and setting of the external parameters may affect the selection and evaluation, the results may not be reliable. Further, the mRMR method considers the average pair-wise correlation amongst features rather than the joint correlation of the selected features.

To measure the scatterness among samples in d-dimensional space, the ratio of between-class scatter matrix and total within-class scatter matrix, has been used successfully for feature selection [12,28] and feature extraction [9,37]. This measure takes the maximum value when between-class scatter matrix and total within-class scatter matrix take maximum and minimum values, respectively, corresponding to a given subset of features. Features in the selected subset distinguish samples of different classes and are not correlated (redundant) with each other. Also, unlike the mRMR method, it does not require setting of the external parameters and discretization of feature values prior to its use.

Motivated by this, we investigated this method to determine a minimal subset of relevant and non-redundant genes in microarray data in this paper. In the original formulation, determining the ratio of scatter matrices takes up huge computation time. Hence, we have proposed an incremental formulation of trace of ratio of scatter matrices (ITRSM) to be used in forward feature selection strategy. The computation time of ITRSM for k features is $O(k^2)$ in comparison to $O(k^3)$ of its batch formulation. To check the efficacy of the proposed method, we have performed extensive experiments on 14 popular, challenging and publicly available datasets. The performance is evaluated in terms of classification accuracy and the number of genes. Experimental results are compared with that of the mRMR method. The Wilcoxon-signed rank test and the Friedman test are carried out to verify the statistical significant difference in the performance of the mRMR and the proposed approach, ITRSM.

The paper is organized as follows. Section 2 briefly discusses the related work for relevant gene selection. Section 3 discusses the mRMR method. The proposed method, ITRSM is presented in Section 4. Experimental setup and results on publicly available microarray datasets are extensively documented in Section 5. The statistical tests are discussed in Section 6. Section 7 includes concluding remarks and future directions.

2. Related work

In literature, many feature selection methods have been proposed for the classification of cancer microarray data by the data mining and pattern recognition community. Since the cancer microarray classification involves high-dimensional feature set, the selection of relevant genes is imperative to improve the performance of the learning system. Many filter, wrapper, and hybrid approaches for feature selection have been used to achieve a smaller set of relevant features.

The univariate filter feature selection approach, Relief [30] and its variant ReliefF [32], information gain [20], and other ranking approaches have been effectively applied for relevant gene selection in cancer microarray classification. The Relief family is claimed to select few attributes effectively. Yang *et al*. [60] proposed two scoring functions based on the means of intraclass

variations and their deviations. Based on these functions, the gene ranking methods GS1 and GS2, which select a stable set of genes without presuming underlying probability distribution, were employed. Although the univariate filter feature selection is fast and scalable, it ignores feature correlation. To overcome this limitation, multivariate filter feature selection approaches are suggested. These methods attempt to select the optimal feature subset rather than identify the individual relevant features. Bo and Jonassen [8] suggested gene pairs-based evaluation method to select gene sets that differentiate between experimental classes. Their work involves the evaluation of *t*-statistics ranking-based gene pairs with four feature subset selection methods namely individual ranking, forward selection, all pairs and greedy pairs with three classifiers. Peng *et al*. [43] suggested the mRMR approach for feature selection which takes care of relevance to class label as well as redundancy in the feature set. But the method considers only pair-wise correlation in features. We will discuss this method in detail in the next section. Zhang *et al*. [62] proposed a two-stage selection algorithm that combines ReliefF to rank the features and mRMR to find a non-redundant subset of genes. Chopra *et al*. [10] proposed the use of gene pairs or doublets from gene regulation pathways to obtain better cancer classification. They have ranked three types of gene pairs using *t*-statistics and evaluated them with four different classifiers. Their study reflects the fact that genes interact to perform a molecular function and the deregulation of pathways is caused by deregulation of interacting genes rather than the individual genes.

The wrapper approach popularized by Kohavi and John [30] searches an optimal feature subset tailored to a particular algorithm and domain. Wrapper methods are computationally intensive but provide better performance. To take advantage of wrapper methods with high-dimensional data, hybrid approaches have been suggested. A well-known hybrid method with recursive elimination of features using the support vector machine (SVM) classifier, support vector machine recursive feature elimination (SVM-RFE) was proposed by Guyon *et al*. [22]. This embedded algorithm discards the attributes which have a small effect on classification, in each iteration. Peng *et al*. [43] used the iterative Genetic Algorithm (GA) with SVM and recursive feature elimination for finding a compact set of predictive genes. GA searches the population in parallel while SVM avoids over-fitting. Goh and Kasabov [18] proposed a hybrid approach that integrates the Pearson correlation coefficient for binning and signal-to-noise ratio for weighted ranking and an adaptive evolving classifier function as a learning model. They have employed the method to bin correlated genes to allow for the selection of the non-redundant genes. The best incremental ranked subset (BIRS) and its variants BIRS_F and BIRS_W [51] attempt to find the best subset of genes based on incremental wrapper-based gene selection. Many research works have investigated bio-inspired computing-based approaches, in conjunction with the wrapper ideology. Hong and Cho [24] and Yu *et al*. [61] used GA for feature selection. Wang *et al*. [58] proposed feature selection using rough sets and particle swarm optimization (PSO).

Recently, Alonso *et al*. [2] have proposed the relaxation of maximum accuracy criteria to select such combination of feature selection and classification algorithm that reduce the number of selected features. Their work involves three feature selection methods: two filter methods and one hybrid i.e. the fast correlation-based filter (FCBF) based on the principle of symmetrical uncertainty, the ReliefF which is a ranking method and the SVM-RFE based on linear SVM, respectively. However, with this relaxation, accuracy is not statistically significantly worse than the best accuracy but the number of genes varies in many cases.

3. Minimum redundancy maximum relevance

In literature many filter methods based on information theory have been suggested by the research community. Mutual information (MI) is an important measure that determines the dependency

amongst variables. Mutual information between two random variables x_i and x_j is defined as

$$I(x_i, x_j) = \iint p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \, \mathrm{d}x_i \, \mathrm{d}x_j,$$

where $p(x_i)$ and $p(x_j)$ are the probability density functions for random variables x_i and x_j, respectively, and $p(x_i, x_j)$ is their joint probability density function. Mutual information between a feature x_i and the target class c is denoted as $I(x_i, c)$. A higher value of mutual information corresponds to more relevance of the feature to the target class. This method is used to select m top-scoring features for classification. But this selected subset of features may contain redundancy due to correlation among the selected subset of genes.

In order to overcome this limitation, Battiti [3] proposed a greedy feature selection method, mutual information-based feature selection (MIFS). MIFS selects a subset of features which maximizes the mutual information about the class, corrected by subtracting a quantity proportional to the average MI with the previously selected features. MIFS handles redundancy at the expense of classification performance. To improve the results, a greedy feature selection method MIFS-U [33] was proposed which gives a better estimate of the mutual information between the input features and the target class as compared to MIFS. The MIFS-U provides the performance of the ideal greedy selection algorithm when the information is distributed uniformly and its computational complexity is almost the same as that of the MIFS. Another improvement over MI has been proposed by Peng *et al*. [43]. They suggested a heuristic framework to minimize redundancy and maximize relevance. In this method, the mutual information between the variables of the selected gene subset needs to be minimized (minimum redundancy criterion), which may be represented as follows:

$$\min_{F \subset G} \frac{1}{|F|^2} \sum_{i,j \in F} I(x_i, x_j),$$

where F is the set of selected features and G is the set of all features.

Also the mutual information between a gene and the target class $c \in \{c_1, c_2, \ldots, c_k\}$ needs to be maximized (maximum relevance criterion) i.e.

$$\max_{F \subset G} \frac{1}{|F|} \sum_{i \in F} I(x_i, c)$$

In order to select a gene subset satisfying both these criterion, the mRMR method is denoted by

$$\max_{F \subset G} \left\{ \sum_{i \in F} I(x_i, c) - \left[\frac{1}{|F|} \sum_{i,j \in F} I(x_i, x_j) \right] \right\}.$$

To use the mRMR method effectively, the probability density function of the data should be known. However, the probability density of the given data is generally not known in advance. Two approaches are often used for approximating the probability density in such cases. One approach involves approximating the probability density using the widely used Parzen window method. This approach, apart from being computationally intensive, depends upon the choice of the window width and the Parzen window function. Another method to approximate the probability density is to discretize the continuous input data. The drawback of this approach is that it is dependent on the choice of the discretization strategy. Moreover, the learning model is developed on the original continuous data whereas the feature selection is based on the discrete data. Apart from the problem of approximating the probability density, the mRMR method described above considers the weighted average of pair-wise correlation instead of considering the joint correlation

among the selected set of features. To find the most relevant set of features, correlation among all selected features needs to be investigated, which mRMR does not consider. Hence, the mRMR method may not select the optimal feature subset in the presence of large number of redundant features. To overcome all the above mentioned limitations, we have investigated a simple and well-known statistical measure which is independent of the external parameters.

4. Incremental formulation of trace of ratio of scatter matrices

In literature, a simple criterion based on the scatter of features in high-dimensional space is recommended, which is the trace of ratio of scatter matrices. The criterion selects those features that are well clustered around their class mean and the features of the different classes are well separated. This criterion function for k-dimensional sample $\mathbf{X}_k = [\mathrm{X}_{.1}, \mathrm{X}_{.2}, \ldots, \mathrm{X}_{.k}]^{\mathrm{T}}$ is given by

$$J_k = \mathrm{trace}(\mathbf{Z}_k),$$

where

$$\mathbf{Z}_k = ((\mathbf{SW}_k)^{-1}\mathbf{SB}_k) \tag{1}$$

$\mathbf{SW}_k$, is the within-class scatter matrix for k-dimensional samples features with m classes, $c_1, c_2, \ldots, c_m$ and is given by

$$\mathbf{SW}_k = \sum_{i=1}^{m} P_i \mathbf{S}_k^i, \tag{2}$$

where $\mathbf{S}_k^i$ is the scatter matrix for k-dimensional samples (k features) of the ith class c_i given by

$$\mathbf{S}_k^i = \sum_{\mathbf{X}_k \in c_i} [(\mathbf{X}_k - \mu_k^i)(\mathbf{X}_k - \mu_k^i)^{\mathbf{T}}]$$

$\mathbf{SB}_k$, the between-class scatter matrix for k-dimensional samples is given by

$$\mathbf{SB}_k = \sum_{i=1}^{m} P_i[(\boldsymbol{\mu}_k^i - \boldsymbol{\mu}_k)(\boldsymbol{\mu}_k^i - \boldsymbol{\mu}_k)^{\mathrm{T}}], \tag{3}$$

where P_i is the a priori probability of the ith class c_i and $\boldsymbol{\mu}_k^i$ and $\boldsymbol{\mu}_k$ are the means of class c_i and whole data for k-dimensional samples, respectively.

Both these matrices $\mathbf{SW}_k$ and $\mathbf{SB}_k$ are symmetric and positive semi-definite. The higher value of criteria J for a particular set of features/genes indicates that the feature set is more discriminatory. Also, the criterion J has the advantage of being invariant under linear transformations.

As a new feature is considered in addition to the existing k features, the new feature vector is represented by $\mathbf{X}_{.k+1} = [\mathrm{X}_{.1}, \mathrm{X}_{.2}, \ldots, \mathrm{X}_{.k+1}]^{\mathrm{T}}$. The corresponding criterion function for the $(k+1)$-dimensional sample vector is given by

$$J_{k+1} = \mathrm{trace}(\mathbf{Z}_{k+1}).$$

The computation of J_{k+1} involves matrix inversion and multiplication of matrices of order $(k+1) \times (k+1)$ and is of time complexity $\mathrm{O}(k^3)$ which is computationally intensive. We now provide a computationally simple, incremental approach to compute the trace($\mathbf{Z}_{k+1}$) in terms of trace($\mathbf{Z}_k$) whose value is already available. The criteria function J_{k+1} can be incrementally computed as follows.

The new mean column vector $\boldsymbol{\mu}^i_{k+1}$ for class c_i and the total mean vector $\boldsymbol{\mu}_{k+1}$ can be incrementally defined as

$$\boldsymbol{\mu}^{\mathbf{i}}_{\mathbf{k+1}} = [(\boldsymbol{\mu}^{\mathbf{i}}_{\mathbf{k}})^{\mathrm{T}} \quad \mu^i_{k+1}]^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{\mu}_{\mathbf{k+1}} = [(\boldsymbol{\mu}_{\mathbf{k}})^{\mathrm{T}} \quad \mu_{k+1}]^{\mathrm{T}},$$

respectively.

Also, the new covariance matrix, $\mathbf{S}^i_{\mathbf{k+1}}$ in terms of $\mathbf{S}^i_{\mathbf{k}}$ can be given by

$$\mathbf{S}^i_{\mathbf{k+1}} = \begin{bmatrix} \mathbf{S}^i_k & \mathbf{b}^i_k \\ (\mathbf{b}^i_k)^{\mathrm{T}} & a^i \end{bmatrix},$$

where $\mathbf{b}^i_k$ is a $k \times 1$ vector and its elements represent covariance $\sigma_{j,k+1}, j = 1, 2, \ldots, k$ and a^i is a scalar covariance element, $\sigma^i_{k+1,k+1}$ of the ith class c_i. The within-class scatter matrix $\mathbf{SW}_{k+1}$ can be written in terms of $\mathbf{SW}_k$ as

$$\mathbf{SW}_{k+1} = \begin{bmatrix} \mathbf{SW}_k & \mathbf{b}_k \\ \mathbf{b}^T_k & a \end{bmatrix},$$

where $\mathbf{b}_k = \sum_{i=1}^m P_i \mathbf{b}^i_k$ is a vector of size $k \times 1$ and $a = \sum_{i=1}^m P_i a^i$ is a scalar.

Using classical results in matrix algebra [19], the inverse of this matrix can be computed as

$$(\mathbf{SW}_{k+1})^{-1} = \begin{bmatrix} (\mathbf{SW}_k - \frac{1}{a}\mathbf{b}_k\mathbf{b}^{\mathrm{T}}_k)^{-1} & -\frac{1}{d}(\mathbf{SW}_k)^{-1}\mathbf{b}_k \\ -\frac{1}{d}\mathbf{b}^T_k(\mathbf{SW}_k)^{-1} & \frac{1}{d} \end{bmatrix}$$

which can also be rewritten as

$$(\mathbf{SW}_{\mathbf{k+1}})^{-\mathbf{1}} = \begin{bmatrix} (\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}} + \frac{1}{d}(\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}}\mathbf{b}_k\mathbf{b}^T_k(\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}} & -\frac{1}{d}(\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}}\mathbf{b}_k \\ -\frac{1}{d}\mathbf{b}^T_k(\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}} & \frac{1}{d} \end{bmatrix},$$

$$(\mathbf{SW}_{\mathbf{k+1}})^{-\mathbf{1}} = \begin{bmatrix} (\mathbf{SW}_{\mathbf{k}})^{-\mathbf{1}} + \frac{1}{d}\mathbf{y}_k\mathbf{y}^T_k & -\frac{1}{d}\mathbf{y}_k \\ -\frac{1}{d}\mathbf{y}^T_k & \frac{1}{d} \end{bmatrix},$$

$$\text{where } \mathbf{y}_k = (\mathbf{SW}_k)^{-1}\mathbf{b}_k \text{ is a vector of order } k \times 1, \tag{4}$$

$$\text{and } d = a - \mathbf{b}^{\mathrm{T}}_k(\mathbf{SW}_k)^{-1}\mathbf{b}_k \text{ is a scalar.}$$

Similarly the between scatter matrix for (k+1)-dimensional sample is given as

$$\mathbf{SB}_{k+1} = \begin{bmatrix} \mathbf{SB}_k & \mathbf{e}_k \\ \mathbf{e}^{\mathrm{T}}_k & f \end{bmatrix},$$

where $\mathbf{e}_k = \sum_{i=1}^m P_i[(\boldsymbol{\mu}^i_k - \boldsymbol{\mu}_k)(\mu^i_{k+1} - \mu_{k+1})]$ is a vector of size $k \times 1$,

$$f = \sum_{i=1}^{m} P_i[(\mu^i_{k+1} - \mu_{k+1})(\mu^i_{k+1} - \mu_{k+1})^{\mathrm{T}}] \text{ is a scalar.}$$

Now $\mathbf{Z}_{k+1}$ can be rewritten as

$$\mathbf{Z}_{k+1} = \begin{bmatrix} \mathbf{Z}_k + \mathbf{y}_k\mathbf{u}_k & (\mathbf{SW}_k)^{-1}\mathbf{e}_k + \frac{1}{d}\mathbf{y}_k v_k \\ -\mathbf{u}_k & -v_k \end{bmatrix}, \tag{5}$$

where

$$\mathbf{u}_k = \frac{1}{q}(\mathbf{b}^{\mathrm{T}}_k\mathbf{Z}_{\mathbf{k}} - \mathbf{e}^{\mathrm{T}}_k) \text{ is a vector of order } 1 \times k, \tag{6}$$

$$v_k = \frac{1}{q}(\mathbf{y}^{\mathrm{T}}_k\mathbf{e}_k - f) \text{ is a scalar.} \tag{7}$$

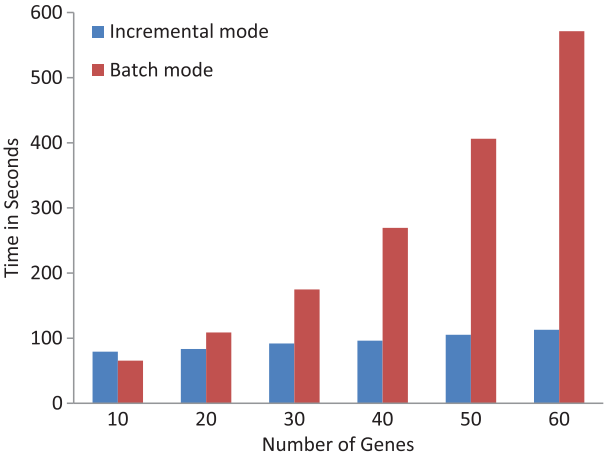

Figure 1. Comparison of computation time of batch and incremental formulation of trace of ratio of scatter matrices.

The criterion measure for $(k+1)$-dimensional sample can now be computed in terms of the available criterion measure of the k-dimensional sample as follows:

$$\text{trace}(\mathbf{Z}_{k+1}) = \text{trace}(\mathbf{Z}_k) + \text{trace}(\mathbf{y}_k \mathbf{u}_k) - v_k. \tag{8}$$

As the terms $\mathbf{y}_k$, $\mathbf{u}_k$ and v_k can be computed in $O(k^2)$, the product of $\mathbf{y}_k$ and $\mathbf{u}_k$ can be computed in $O(k)$. Since the term $\mathbf{Z}_k$ (Equation 1) is already available from the last iteration, the incremental evaluation of $\mathbf{Z}_{k+1}$ can be done in $O(k^2)$. Therefore, with the proposed incremental computation, criterion J requires the quadratic-order computations.

For demonstration, the computational time variation of batch and incremental mode, for the selection of k features ($k = 10, 20, 30, 40, 50, 60$) from the Prostate dataset is shown in Figure 1. It can be observed from Figure 1 that for fewer genes, batch mode takes less time than the incremental mode. However, as the number of genes increases, the computational time for the batch mode is significantly more in comparison to the incremental mode.

Algorithm: Incremental Formulation of Trace of Ratio of Scatter Matrices (ITRSM)
Input-F (Set of all genes), Num_genes (required number of genes)
Output-S (Set of selected genes)

(1) Initialization: Set F =' Set of all genes', S = Φ (set of selected genes), $k = 0$.
(2) For each gene x_i in F
  i. Calculate $\mathbf{SW}_1$ and $\mathbf{SB}_1$ according to eqs(2 and 3) for $k = 1$.
  ii. Calculate $\mathbf{Z}_1 = ((\mathbf{SW}_1)^{-1}\mathbf{SB}_1)$.
  iii. Calculate $J_1 = \text{trace}(\mathbf{Z}_1)$ for each gene.
(3) Select the gene x_h such that $x_h = \arg\max_{x_i \in F} J_1$
(4) $k = k + 1$.
(5) $S = S \cup \{x_h\}$, F=F $- \{x_h\}$
(6) While k <=Num_genes
  (a) For each gene x_j in F
    i. Consider set T $= S \cup \{x_j\}$
    ii. Using the set T compute $\mathbf{Z}_{k+1}$ incrementally using eq. (5) i.e.

$$\mathbf{Z}_{k+1} = \begin{bmatrix} \mathbf{Z}_k + \mathbf{y}_k \mathbf{u}_k & (\mathbf{SW}_k)^{-1}\mathbf{e}_k + \frac{1}{d}\mathbf{y}_k v_k \\ -\mathbf{u}_k & -v_k \end{bmatrix}$$

Where $\mathbf{y}_k$, $\mathbf{u}_k$ and v_k are calculated using Equations (4), (6), (7)

iii. Calculate J_{k+1} i.e. $trace(\mathbf{Z}_{k+1})$ using Equation (8)

$$trace\,(\mathbf{Z}_{k+1}) = trace\,(\mathbf{Z}_k) + trace\left(\mathbf{y}_k\mathbf{u}_k\right) - v_k$$

(b) Select the gene x_l such that $x_l = \arg\max_{x_j \in F} J_{k+1}$

(c) $S = S \cup \{x_l\}$, F=F $-$ $\{x_l\}$

(d) $k = k$+1

5. Experimental setup and results

To check the efficacy of the proposed approach ITRSM, we have done extensive experiments on 14 well-known and publicly available cancer microarray datasets. The performance is evaluated in terms of the classification accuracy and the number of relevant genes. The performance of the proposed method is compared with that of the mRMR method. While comparing the two methods, the one with the higher classification accuracy is termed as the better performer while in the case of the same classification accuracy, the method with lesser number of selected genes is considered better. Three popular classifiers viz. K-nearest neighbour (KNN), linear discriminant classifier (LDC) and SVM, commonly used by machine learning and data mining communities, are used for evaluation.

A brief description of the datasets used in our experiment is given in Table 1. We applied the below mentioned preprocessing procedures on the datasets. The endometrium and the breast datasets contained null values. For these two datasets, attributes with more than 30% of missing values have been discarded [40,45]. The remaining attributes with missing values were replaced with their respective class-wise mean. For the NCI60 dataset, a class having only two samples has been removed from the dataset during the experiment. Also in NCI60, since the number of samples belonging to each class is very small, 2000 genes with the highest variance were considered. Similarly for Leukemia_c, endometrium and breast datasets, 3000, 3000 and 5000 genes, respectively, with the highest variance were selected. For the preprocessing of the remaining datasets (Table 2), a preprocessing strategy similar to that given by Ramaswamy *et al*. [48] and Yang *et al*. [60] is followed. Two cut-off values: Floor and Ceil were used. For each dataset, the expression levels below Floor and above Ceil were set to Floor and Ceil, respectively. Max/Min

Table 1. Description of datasets.

S. no.	Dataset	Samples	Original genes	Preprocessed genes	Classes
1.	Prostate [54]	102	12600	5966	2
2.	CNS-v1 [47]	34	7129	2277	2
3.	Colon [1]	62	2000	2000	2
4.	Leukemia_c [11]	28	12625	3000	2
5.	Breast [57]	97	24482	5000	2
6.	Colon_l [34]	37	22883	8826	2
7.	Endometrium [49]	42	8872	3000	4
8.	NCI 60 [50]	60	9706	2000	9
9.	CNS-v2 [47]	40	7129	5548	5
10.	Glioma [41]	50	12625	4434	4
11.	SRBCT [29]	83	2308	2308	4
12.	Melanoma [7]	38	8067	8067	3
13.	Leukaemia [20]	72	7129	7129	3
14.	GCM [48]	198	16063	11328	14

Table 2. Preprocessing strategy.

Dataset name	Floor	Ceiling	Max/min	Max–min
Prostate	100	16,000	5	50
CNS-v1	20	16,000	5	500
CNS-v2	20	16,000	5	500
Colon-l	20	16,000	3	100
Glioma	20	16,000	3	100
GCM	20	16,000	5	500

ratio and Max–Min difference of a gene across samples was used to filter the genes with little variation across the samples. Preprocessing was followed by normalization using z-score.

The training and the test data of GCM is separately available. Hence, the classification accuracy of GCM dataset is reported using the test data. The classification accuracy of remaining datasets is given in terms of leave one out cross-validation (LOOCV), 10-fold cross-validation and 5-fold cross-validation. For 10-fold and 5-fold CV, the average classification accuracy of 50 runs is reported.

For KNN, optimal value of k is chosen with respect to the leave-one-out accuracy on the dataset. In the case of SVM, linear kernel is used with grid search optimization over the range $C = 10^0, 10^1, 10^2, 10^3$. The experiments are conducted using Matlab.

For mRMR, the data were discretized into a binary set of values and two different types of ternary values as suggested by Peng *et al*. [43]. Then, a set of 60 top-ranked genes was extracted for each of the 14 datasets. Similarly, a set of 60 top-ranked genes was selected using the proposed method (ITRSM). For both these methods, the top-ranked genes were incrementally included one by one to develop the decision model. At every stage, the classification accuracy of the test data was determined. The maximum classification accuracy obtained for each combination of filter method and a classifier for a given dataset is shown in Table 3. The best classification accuracy achieved for each dataset is shown in bold. The number within parenthesis represents the number of genes corresponding to which maximum classification accuracy is obtained for a given filter method and a classifier. Also, we have listed the gene accession numbers for the best results i.e. maximum accuracy and minimum number of genes combination, in Table 4.

In literature, various gene selection methods have been suggested and applied on these datasets. In Table 5, a comparison of classification accuracy and number of genes, of the proposed method (ITRSM) with existing gene selection methods is listed.

The following observations can be made from Tables 3 to 5:

- For all the two class datasets, the proposed method (ITRSM) outperforms mRMR in terms of the classification accuracy for all the three classifiers using LOOCV, 10-fold and 5-fold cross-validation except in the case of the breast dataset for 5-fold and 10-fold cross validations.
- The proposed method (ITRSM) also outperforms mRMR in terms of the number of selected genes for most of the two class datasets with all the three classifiers.
- For all the multiclass datasets, the proposed method (ITRSM) provides better or same classification accuracy in comparison to mRMR with LDC using LOOCV, 10-fold and 5-fold cross-validation, except for the SRBCT dataset.
- For most of the multiclass datasets, mRMR performs better in comparison to the proposed method (ITRSM) with KNN and SVM.
- For most of the datasets, the performance of the proposed method (ITRSM) in combination with LDC is better using LOOCV, 10-fold and 5-fold cross-validations.

Table 3. Comparison of ITRSM and mRMR methods in terms of classification accuracy and number of genes.

Dataset	No. of classes	Classifier	LOOCV		10fold		5fold	
			ITRSM	mRMR	ITRSM	mRMR	ITRSM	mRMR
Prostate	2	KNN	**98.04(5)**	97.06(10)	**97.35(5)**	96.71(10)	**96.88(5)**	96.35(10)
		LDC	**100(10)**	96.08(4)	**100(11)**	95.65(8)	**100(17)**	95.49(8)
		SVM	**100(7)**	97.06(8)	**100(11)**	97.06(5)	**100(7)**	97.06(5)
CNS-v1	2	KNN	**100(4)**	**100(4)**	**100(10)**	96.76(27)	**99.82(9)**	95.71(27)
		LDC	**100(4)**	97.05(9)	**100(7)**	95.41(50)	**100(11)**	93.47(54)
		SVM	**100(4)**	97.06(2)	**100(3)**	94.12(4)	**100(14)**	94.12(4)
Colon	2	KNN	**95.16(9)**	90.32(10)	**94.39(9)**	88.19(11)	**92.48(9)**	87.52(52)
		LDC	**100(21)**	88.70(8)	**100(22)**	88.39(8)	**100(29)**	87.45(8)
		SVM	**98.39(28)**	90.32(12)	**98.39(13)**	90.32(14)	**98.39(13)**	88.71(10)
Leukemia_c	2	KNN	**92.86(4)**	85.71(28)	**89.21(4)**	85.07(28)	**87.21(4)**	82.86(28)
		LDC	**100(5)**	96.42(29)	**100(5)**	85.79(46)	**100(5)**	81.64(45)
		SVM	**100(8)**	85.71(11)	**96.43(5)**	89.29(39)	**100(6)**	89.29(43)
Breast	2	KNN	**86.6(30)**	83.51(31)	63.48(24)	**69.07(27)**	61.2(25)	**66.78(27)**
		LDC	**100(19)**	77.32(22)	64.31(6)	**68.16(47)**	64.19(6)	**65.61(47)**
		SVM	**100(17)**	84.54(56)	67.01(6)	**75.26(42)**	67.01(6)	**79.38(53)**
Colon_l	2	KNN	**100(3)**	**100(3)**	99.78(10)	100(5)	99.78(10)	**100(5)**
		LDC	**100(4)**	**100(4)**	**100(4)**	100(7)	**100(5)**	100(7)
		SVM	**100(4)**	100(6)	**100(1)**	100(1)	**100(1)**	100(1)
Nci60	8	KNN	74.14(58)	**82.76(26)**	72.21(59)	**77.45(38)**	74.52(58)	**78.14(37)**
		LDC	**100(39)**	82.76(57)	**96.03(37)**	74.59(59)	**79.93(27)**	70.48(59)
		SVM	75.86(60)	**82.76(45)**	75.86(52)	**81.03(54)**	70.69(60)	**77.59(13)**
CNS- v2	5	KNN	85.00(14)	**97.5(14)**	83.3(14)	**94.55(15)**	82.5(14))	92.5(14)
		LDC	**100(17)**	95.00(16)	**97.5(17)**	95.00(16)	**95.00(19)**	92.5(19)
		SVM	77.5(10)	**97.5(8)**	82.5(10)	**97.5(8)**	80.00(16)	**97.5(9)**
Glioma	4	KNN	82.00(41)	**92.00(45)**	80.2(45)	**89.52(51)**	79.84(44)	**89.16(48)**
		LDC	**100(23)**	90.00(7)	**100(24)**	86.52(7)	**100(25)**	85.8(5)
		SVM	84.00(28)	**92.00(14)**	86.00(28)	**88(11)**	84.00(28)	**88(5)**
Endometrium	4	KNN	92.86(23)	**95.24(14)**	92.86(25)	**95.24(10)**	92.86(25)	**95.24(10)**
		LDC	**100(20)**	95.24(21)	**97.82(21)**	92.86(21)	**95.24(21)**	92.86(25)
		SVM	92.86(52)	**95.24(14)**	92.86(50)	**95.24(10)**	92.86(50)	**95.24(10)**
SRBCT	4	KNN	93.98(44)	**100(10)**	90.96(60)	**100(20)**	90.17(60)	**100(20)**
		LDC	100(55)	**100(20)**	98.87(57)	**99.64(32)**	94.87(41)	**98.70(32)**
		SVM	100(53)	**100(17)**	97.60(38)	**100(20)**	97.60(40)	**100(20)**
Melanoma	3	KNN	**84.21(3)**	78.95(22)	**83.21(3)**	75.26(48)	**81.47(3)**	76.26(49)
		LDC	**89.47(12)**	78.95(9)	**85.95(15)**	70.42(6)	**82.68(10)**	69.95(6)
		SVM	**84.21 (5)**	84.21(29)	**86.84(3)**	86.84(14)	**86.84(15)**	78.95(13)
Leukaemia	3	KNN	97.22(51)	**98.61(16)**	95.44(60)	**98.56(31)**	94.11(60)	**98.28(45)**
		LDC	**100(44)**	97.22(5)	**98.67(50)**	96.92(19)	93.72(47)	**96.58(19)**
		SVM	**100(55)**	98.61(18)	**100(55)**	98.61(28)	98.61(55)	**98.61(14)**
GCM	14	KNN	52.17(43)	**52.17(29)**	Train and test observations			
		LDC	**50.00(17)**	50.00(32)				
		SVM	**44.44(17)**	42.6(20)				

- Mostly, the classification accuracy of both the methods is less for a smaller value of K in comparison to the classification accuracy achieved with a higher value of K, using K-fold cross validation.
- It can be observed from Table 5 that the performance of the proposed method (ITRSM) is better or comparable to the existing feature selection methods in terms of classification accuracy except in the case of GCM. Since the GCM dataset is a challenging classification problem, most filter methods do not perform well. The results of GCM with the proposed ITRSM approach are comparable to those of mRMR. However, computationally intensive wrapper methods give better results in most cases.

Table 4. Accession numbers of microarray datasets corresponding to maximum classification accuracy.

Dataset	Method	Best accuracy	Gene list
Prostate	ITRSM + SVM	100(10)	37639_at 41468_at 36554_at 41504_s_at 34533_at 35304_at 32137_at
CNS-v1	ITRSM + LDC	100(4)	HG3543-HT3739_atU25789_at X04741_at X71973_at
Colon	ITRSM + LDC	100(21)	M63391 D00860 R80427 H55916 R88740 M26683 M90516 T64012 T72863 K02268 R34876 L41142 M31303 L05144 X52541 H02465 H70425 H79575 L10284 R70030 M63239
Leukemia_c	ITRSM + LDC	100(5)	36780_at 32178_r_at 41166_at 39302_at 35271_at
Breast	ITRSM + SVM	100(17)	D42044 Contig28522_RC Contig51517_RC Contig67229_RC AL133052 U03886 NM_003079 NM_001756 NM_000909 NM_001799 Contig48008_RC Contig37897_RC AB033085 Contig36859_RC Contig30573_RC Contig42162_RC AF059531
Colon_l	ITRSM + KNN	100(3)	205191_at 203067_at 206492_at
Nci60 (indices)	ITRSM+LDC	100(39)	581 276 740 935 932 426 979 424 463 250 783 63 84 490 863 866 469 578 168 765 920 987 874 473 158 771 534 704 361 668 5 933 797 311 550 800 555 724 660
CNS-v2	ITRSM + LDC	100(17)	D87463_at M92449_at M69197_xpt2_s_at D21267_at Z31695_at X05309_at HG2259-HT2348_s_at U07358_at D83735_at D78611_at U70671_at M30838_at U89336_cds8_at S56151_s_at U36341_rna1_at AF002700_at M96995_s_at
Glioma	ITRSM + LDC	100(23)	35905_s_at 1030_s_at 39556_at 41288_at 1367_f_at 38020_at 428_s_at 40359_at 7697_s_at 33879_at 41233_at 33433_at 34861_at 1875_f_at 37722_s_at 1775_at 37001_at 41280_r_at 37389_at 40160_at 34359_at 36261_at 39342_at
Endometrium	ITRSM + LDC	100(20)	166898 162203 160361 164942 167068 161948 165246 163316 161692 165437 160273 162551 163558 165826 162870 164531 163270 168067 169025 163826
SRBCT	mRMR + KNN	100(10)	207274 770394 812105 796258 377461 530185 814526 784224 1435862 796475
Melanoma	ITRSM + LDC	89.47(12)	898258 810062 128100 563403 666639 258061 470379 130044 898092 34327 230976 810934
Leukaemia	ITRSM + LDC	100(44)	X03934_at X66533_at M21624_at X00437_s_at M28826_at M86699_at M37271_s_at S34389_at X66945_at X15357_at D31883_at L08177_at X87241_at U91316_at S78187_at Z72499_at M32304_s_at D87116_at U76638_at L10838_at D21163_at U20647_at U10868_at X69433_at D00760_at D83702_at U02031_at U49869_rna1_at U91985_at M85169_at HG2562-HT2658_s_at HG1595-HT4788_s_at L01087_at X92106_at AFFX-HUMGAPDH/M33197_M_at M60278_at D14689_at D82348_at Y07755_at U43185_s_at M24486_s_at U03911_at U07132_at X55740_at
GCM	mRMR + KNN	52.17(29)	AFFX-HSAC07/X00351_5_at K03460_at L15409_at D90276_at L41067_at L10838_at AFFX-HUMGAPDH/M33197_5_at M12529_at D30758_at D13988_at M73720_at L42451_at HG2463-HT2559_at AFFX-HSAC07/X00351_M_at HG1800-HT1823_at M83181_at D86964_at HG2810-HT2921_atM33518_at D28416_at L19161_at D87024_at D38076_at M37766_at HG1602-HT1602_at J03798_at D83597_at M60091_at M63379_at

Table 5. Comparison of ITRSM with the state-of-art methods.

Prostate		CNS-v1		Breast		Colon	
ITRSM + SVM	100(7)	ITRSM + LDC	100(4)	ITRSM + SVM	100(17)	ITRSM + LDC	100(21)
GAKNN [35]	96.3(79)	Alonso(SVM- RFE + SVM) [2]	75.49(100)	Alonso *et al.* [2]	70.96(5)	PSO + ANN [56]	88.7
BIRS [51]	91.2(3)	TSP [10]	77.9	IFS [18]	94.9(9)	WFFSA-G [63]	97.9(100)
Hong and Cho [24]	96.3(79)	k-TSP [10]	97.1	Correlation with output [18]	89.7(105)	BIRSW [51]	85.48
		PAM [10]	82.35	SNR [18]	91(81)	BIRSF [51]	85.48
						Pirooznia *et al.* [46]	95.56
Colon_l		Leukemia_c		NCI60		CNS-v2	
ITRSM + LDC	100(3)	ITRSM + LDC	100(5)	ITRSM + LDC	100(39)	ITRSM + LDC	100(17)
Irgon *et al.* [25]	96.41	Alonso *et al.* [2]	78.22(100)	mRMR + Relief-F + SVM [62]	68.33	OVO SVM [61]	84.4
				WFFSA-G [63]	85.25(14)	OVRSVM [61]	85.6
				GA/MLHD [55]	85.37(13)	Yu *et al.* [61]	87.8
				GA/SVM/RFE [44]	87.93(27)		
Endometrium		SRBCT		Glioma		Melanoma	
ITRSM+LDC	**100(20)**	ITRSM + SVM	100(53)	**ITRSM+LDC**	**100(23)**	ITRSM + LDC	89.47(12)
DICLENS [38]	100	GS2 + SVM [61]	100(96)	Shirahata *et al.* [53]	96.6(67)	Pirooznia *et al.* [46]	94.74
MCLA [38]	92	Gs1 + SVM [61]	98.8(34)	Alonso (relief + 3-NN) [2]	55.56(25)	DICLENS [38]	89
CSPA [38]	55	Ftest + SVM [61]	100(78)	Yu *et al.* [61]	78	MCLA [38]	89
HGPA [38]	42	Tibsirani [56]	100(43)	OVO SVM [61]	72	HGPA [38]	70
LCE [38]	92			OVR SVM [61]	78	CSPA [38]	89
Leukaemia		GCM					
ITRSM + LDC	100(44)	ITRSM + KNN	52.17(43)				
GS2 + KNN [60]	98.6(10)	mRMR + Relief-F + SVM [62]	64.65				
GS1 + SVM [60]	98.6(4)	Wang *et al.* [58]	69.8(28)				
Ftest + SVM [60]	98.6(33)	GA/MLHD [54]	79.33(32)				
Tibsirani [56]	100(21)	OVA-SVM [48]	78				
		GA/SVM/RFE [44]	85.19(26)				

Table 6. Ranking of difference of classification accuracy of mRMR (d_1) and ITRSM (d_2) according to Wilcoxon-signed rank test.

Dataset	Classifier	$d_i = d_2 - d_1$	Rank
Prostate	KNN	0.98	10
	LDC	3.92	22
	SVM	2.94	17.5
Cns-v1	KNN	0	5
	LDC	2.95	19
	SVM	2.94	17.5
Colon	KNN	4.84	24
	LDC	11.3	36
	SVM	8.07	31
Leukemia_c	KNN	7.15	29
	LDC	3.58	21
	SVM	14.29	38
Breast	KNN	3.09	20
	LDC	22.68	42
	SVM	15.46	39
Colon-l	KNN	0	5
	LDC	0	5
	SVM	0	5
Nci-60	KNN	−8.62	32
	LDC	17.24	40
	SVM	−6.9	28
Cns-v2	KNN	−12.5	37
	LDC	5	25
	SVM	−20	41
Endometrium	KNN	−2.38	14.5
	LDC	4.76	23
	SVM	−2.38	14.5
SRBCT	KNN	−6.02	27
	LDC	0	5
	SVM	0	5
Melanoma	KNN	5.26	26
	LDC	10.52	35
	SVM	0	5
Leukaemia	KNN	−1.39	11.5
	LDC	2.78	16
	SVM	1.39	11.5
Glioma	KNN	−10	33.5
	LDC	10	33.5
	SVM	−8	30
GCM	KNN	0	5
	LDC	0	5
	SVM	1.84	13

6. Statistical analysis of result

To determine whether the two feature selection methods are significantly different or not, a statistical comparison is generally conducted. Statisticians have suggested that non-parametric tests are more suitable when data distribution assumptions are not valid. Hence, the Wilcoxon-signed rank test [59] and the Friedman test [15,16] are applied for hypothesis testing.

The Wilcoxon-signed rank test [13,59] is analogous to the paired *t*-test in nonparametric statistical procedures. It is a pair-wise test that aims to detect the significant differences between two sample means. In our case the null hypothesis, H_0 is the equivalence of the two filter methods and the alternative hypothesis, H_1 indicates significant difference between the two. To apply the Wilcoxon-signed rank test, firstly the difference d_i between the performance scores of the two methods (mRMR and ITRSM) for all datasets with the three classifiers is obtained. Then the

Table 7. Ranking of mRMR and ITRSM methods' ranks according to Friedman test.

Dataset	Classifier	ITRSM ranks	mRMR ranks
Prostate	KNN	1	2
	LDC	1	2
	SVM	1	2
Cns-v1	KNN	1.5	1.5
	LDC	1	2
	SVM	1	2
Colon	KNN	1	2
	LDC	1	2
	SVM	1	2
Leukemia_c	KNN	1	2
	LDC	1	2
	SVM	1	2
Breast	KNN	1	2
	LDC	1	2
	SVM	1	2
Colon-l	KNN	1.5	1.5
	LDC	1.5	1.5
	SVM	1	2
Nci-60	KNN	2	1
	LDC	1	2
	SVM	2	1
Endometrium	KNN	2	1
	LDC	1	2
	SVM	2	1
SRBCT	KNN	2	1
	LDC	1	2
	SVM	2	1
Melanoma	KNN	2	1
	LDC	1.5	1.5
	SVM	1.5	1.5
Leukaemia	KNN	1	2
	LDC	1	2
	SVM	1.5	1.5
Glioma	KNN	1.5	1.5
	LDC	1.5	1.5
	SVM	1.5	1.5
Cns-v2	KNN	2	1
	LDC	1	2
	SVM	2	1
GCM	KNN	2	1
	LDC	1	2
	SVM	1	2

differences are ranked (Table 6) according to their absolute values; average ranks are assigned in the case of ties. Let $R+$ be the sum of ranks for the datasets on which the second algorithm outperformed the first, and $R-$ the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored.

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i) \quad R^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i).$$

The sum of the ranks for the positive difference R^+ and the negative difference R^- is 609 and 289, respectively. W, the smaller of the sums, is given by $\min(R+, R-)$. For our experiments, W is 289.

The z-statistic is defined as [13,59]

$$z = \frac{W - (1/4)N(N+1)}{\sqrt{(1/24)N(N+1)(2N+1)}}.$$

For our datasets, the value of z is -2.03176 whose corresponding p-value is 0.0422. This suggests that the proposed method, ITRSM outperforms mRMR with 95% confidence level.

The Friedman test [15] is another nonparametric statistical test that is equivalent of the parametric analysis of variance. It ranks the algorithms for each dataset separately, the best performing algorithm is assigned the rank value 1, and the second best is given rank 2 as shown in Table 7. In the case of ties, average ranks are assigned. For each of the ranking algorithms, mRMR and the proposed method (ITRSM), the average rank is measured. The Friedman test checks whether the measured average ranks are significantly different from the mean rank R_j as expected under the null-hypothesis.

Let r_i^j be the rank of the jth of k algorithms on the ith of N datasets. The Friedman test compares the average ranks of algorithms, $R_j = (1/N)\sum_i r_i^j$. The Friedman statistic is calculated as

$$\chi_F^2 = \frac{12N}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right],$$

which is distributed according to χ_F^2 with $k-1$ degrees of freedom.

For our experimental results, R_1 and R_2 take values 1.345238 and 1.654762, respectively. With the p-value as 0.044 the hypothesis is rejected with a 95% confidence factor, hence establishing that the accuracy of the proposed method (ITRSM) is significantly different from that of mRMR. Thus both the tests rejected the null hypothesis, which means that the two methods differ significantly in their performance and the proposed method (ITRSM) performs better.

7. Conclusion

Microarray gene expression data are often characterized by high dimension and a small sample size. Hence microarray cancer classification suffers from the curse of dimensionality. To overcome this limitation, feature selection methods have been applied in literature. These determine a minimal set of relevant features that help in improving classification accuracy and alleviating the computational burden. In this paper, we proposed the ITRSM which is used to determine a relevant set of genes. Unlike the mRMR method, the proposed method does not require any discretization or external parameter setting. The time complexity of proposed feature selection method is better in comparison to its batch formulation. Further, the computation involved is independent of the external parameters. To check the efficacy of the proposed method, we have carried out extensive experiments on 14 publicly available cancer microarray datasets. Experimental results show that the proposed method outperforms the well-known mRMR method and the other existing feature selection methods. Two nonparametric tests, Wilcoxon-signed rank and Friedman test show that the two methods differ significantly in their performance and the proposed method (ITRSM) performs better in comparison to mRMR.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive suggestions which significantly helped in improving the quality of this paper.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligo-nucleotide arrays*, Proc. Natl. Acad. Sci. USA 96(12) (1999), pp. 6745–6750.

[2] C.J. Alonso-Gonzalez, Q.I. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, *Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods*, Exp. Syst. Appl. 39 (2012), pp. 7270–7280.

[3] R. Battiti, *Using mutual information for selecting features in supervised neural net learning*, IEEE Trans. Neural Netw. 5 (1994), pp. 537–550.

[4] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, 1961.

[5] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, *Tissue classification with gene expression profiles*, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, ACM Press, New York, 2000, pp. 54–64.

[6] C. Bhattacharya, L.R. Grate, A. Rizki, D. Radisky, F.J. Molina, M.I. Jordan, M.J. Bissell, and I.S. Mian, *Simultaneously classification and relevant feature identification in high dimensional spaces: application to molecular profiling data*, Signal Process. 83(4) (2003), pp. 729–743.

[7] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, *Molecular classification of cutaneous malignant melanoma by gene expression profiling*, Nature 406 (6795) (2000), pp. 536–540.

[8] T. Bo and I. Jonassen, *New feature subset selection procedures for classification of expression profiles*, Genome Biol. 3 (2002), pp. 1–0017.

[9] L. Chao, J. Xiaoyuan, Z. David, G. Shiqiang, and Y. Jingyu, *Discriminant subclass-center manifold preserving projection for face feature extraction*, IEEE International Conference on Image Processing, Brussels, Belgium, IEEE, Piscataway, NJ, 2011, pp. 3013–3016.

[10] P. Chopra, J. Lee, J. Kang, and S. Lee, *Improving cancer classification accuracy using gene pairs*, PloS ONE 5(12) (2010), pp. e14305.

[11] L.C. Crossman, M. Mori, Y.C. Hsieh, T. Lange, P. Paschka, and C.A. Harrington, *Chronic myeloid leukemia white cells from cytogenetic responders and non responders to imatinib have very similar gene expression signatures*, Haematologica 90 (2005), pp. 459–464.

[12] T.H. Dat and C. Guan, *Feature selection based on fisher ratio and mutual information analyses for robust brain computer interface*, IEEE ICASSP, Honolulu, Hawai, USA, 2007, pp. I–337.

[13] J. Demsar, *Statistical comparisons of classifiers over multiple data sets*, J. Mach. Learn. Res. 7 (2006), pp. 1–30.

[14] R. Duda, P. Hart, and D. Stork, *Pattern classification*, 2nd ed., Wiley Interscience Publication, Chichester, 2000.

[15] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, J. Amer. Statist. Assoc. 32 (1937), pp. 675–701.

[16] M. Friedman, *A comparison of alternative tests of significance for the problem of m rankings*, Ann. Math. Statist. (1940), pp. 1186–1192.

[17] L.M. Fu and C.S.F. Liu, *Evaluation of gene importance in microarray data based upon probability of selection*, BMC Bioinformatics 6(1) (2005), pp. 1–11.

[18] H.L. Goh and N. Kasabov, *An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data*, J. Bioinformatics Comput. Biol. 3(5) (2005), pp. 1107–1136.

[19] G.H. Golub and C.F.V. Loan, *Matrix Computations*, 3rd ed., John Hopkins University Press, Baltimore, MD, 1996.

[20] T.R. Golub, D.R. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science 286 (1999), pp. 531–537.

[21] I. Guyon and A. Elisseff, *An introduction to variable and feature selection*, Mach. Learn. Res. 3 (2003), pp. 1157–1182.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machine*, Mach. Learn. 46(1–3) (2002), pp. 389–422.

[23] A. Haury, P. Gestruad, and J. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures*, PLoS ONE 6(12) (2011).

[24] J.H. Hong and S.B. Cho, *The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming*, Artif. Intell. Med. 36 (2006), pp. 43–58.

[25] J. Irgon, C. Huang, Y. Zhang, D. Talantov, G. Bhanot, and S. Szalma, *Robust multi-tissue gene panel for cancer detection*, BMC Cancer 10 (2010), pp. 319.

[26] A. Jain and D. Zongker, *Feature selection, evaluation, application and small sample performance*, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997), pp. 153–158.

[27] Z. Ji, P. Jing, Y. Su, and Y.W. Pang, *Rank canonical correlation analysis and its application in visual search reranking*, Signal Process. 93 (2013), pp. 2352–2360.

[28] H. Ketabdar, J. Richiardi, and A. Drygajlo, *Global feature selection for on-line signature verification*, Proceedings of the 12th International Graphonomics Society Conference, Salerno, Italy, 2005.

[29] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann , F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Nat. Med. 76 (2001), pp. 673–679.
[30] K. Kira and L.A. Rendell, *The feature selection problem: Traditional methods and a new algorithm*, AAAI-92 Proceedings, San Jose, CA, 1992, pp. 129–134.
[31] R. Kohavi and G. John, *Wrapper for feature subset selection*, Artif. Intell. 97(1–2) (1997), pp. 273–324.
[32] I. Kononenko, *Estimating attributes: Analysis and extensions of RELIEF*, *European Conference on Machine Learning*, Catania, Italy, Springer, Berlin, 1994, pp. 171–182.
[33] N. Kwak and C.H. Choi, *Input feature selection for classification problems*, IEEE Trans. Neural Netw. 131 (2002), pp. 143–159.
[34] P. Laiho, A. Kokko, S. Vanharanta, R. Salovaara, H. Sammalkorpi, H. Jarvinen, J.P. Mecklin, T.J. Karttunen, K. Tuppurainen, V. Davalos, S. Schwartz, D. Arango, M.J. Makinen, and L.A. Aaltonen, *Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis*, Oncogene 26(2) (2007), pp. 312–320.
[35] L. Li, C.R. Weinberg, T.A. Darden, L.G. Pedersen, *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) (2001), pp. 1131–1142.
[36] T. Li, C. Zhang, and M. Ogihara, *Comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression*, Bioinformatics 20 (2004), pp. 2429–2437.
[37] Z. Liu, J. Huang, and Y. Wang, *Audio feature extraction and analysis for scene classification*, Proc. IEEE 1st Multimedia Workshop, Princeton, NJ, IEEE, Piscataway, NJ, 1997, pp. 343–348.
[38] S. Mimaroglu and E. Aksehirli, *DICLENS: Divisive clustering ensemble with automatic cluster number*, IEEE/ACM Trans. Comput. Biol. Bioinformatics 9(2) (2012), pp. 408–420.
[39] T.M. Mitchell, *The need for biases in learning generalizations*, Tech. Rep. CBM-TR-117, Rutgers University, New Brunswick, NJ, 1980.
[40] M. Motoori, I. Takemasa, M. Yano, S. Saito, H. Miyata, S. Takiguchi, Y. Fujiwara, T. Yasuda, Y. Doki, Y. Kurokawa, N. Ueno, S. Oba, S. Ishii, M. Monden, and K. Kato, *Prediction of recurrence in advanced gastric cancer patients after curative resection by gene expression profiling*, Int. J. Cancer 114 (2005), pp. 963–968.
[41] C.L. Nutt, D.R. Mani, R.A. Betensky, P. Tamayo, J.G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M.E. McLaughlin, T.T. Batchelor, P.M. Black, A. von Deimling, S.L. Pomeroy, T.R. Golub, and D.N. Louis, *Gene expression based classification of malignant gliomas correlates better with survival than histological classification*, Cancer Res. 63(7) (2003), pp. 1602–1607.
[42] Y.W. Pang, H. Yan, Y. Yuan, and K. Wang, *Robust cohog feature extraction in human centered image/video management system*, IEEE Trans. Systems, Man, Cybern. Part B: Cybernetics 42 (2012), pp. 1–11.
[43] H. Peng, F. Long, and C. Ding, *Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy*, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005), pp. 1226–1238.
[44] S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, and L. Chen, *Molecular classification of cancer types from microarray data using combination of genetic algorithm and support vector machine*, Science Direct, FEBS Lett. 555(2) (2003), pp. 358–362.
[45] J.Y. Pierga, J. Reis-Filho, S.J. Cleator, T. Dexter, A. Mackay, P. Simpson, and K. Fenwick, *Microarray-based comparative genomic hybridization of breast cancer patients receiving neoadjuvant chemotherapy*, Br. J. Cancer 96(2) (2006), pp. 341–351.
[46] M. Pirooznia, J. Yang, M.Q. Yang, and Y. Deng, *A comparative study of different machine learning methods on microarray gene expression data*, BMC Genom. 9(Suppl. 1) (2008), S13.
[47] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, *Prediction of central nervous system embryonal tumour outcome based on gene expression*, Nature 415(6870) (2002), pp. 436–442.
[48] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, *Multiclass cancer diagnosis using tumor gene expression signatures*, Proc. Natl. Acad. Sci. USA 98(26) (2001), pp. 15149–15154.
[49] J.I. Risinger, G.L. Maxwell, G.V.R. Chandramouli, A. Jazaeri, O. Aprelikova, T. Patterson, A. Berchuck, and J.C. Barrett, *Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer*, Cancer Res. 63 (2003), pp. 6–11.
[50] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van De Rijn, M. Walthamet, A. Pergamenschikov, J. C. Lee, D. Lashkari , D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, *Systematic variation in gene expression patterns in human cancer cell lines*, Nat. Genet. 24 (2000), pp. 227–235.
[51] R. Ruiz, J.C. Riqueline, and J.S. Aguilar-Ruiz, *Incremental wrapper based gene selection from microarray data for cancer classification*, Pattern Recogn. 39(12) (2006), pp. 2383–2392.
[52] S. Shah and A. Kusiak, *Cancer gene search with data mining and genetic algorithms*, Comput. Biol. Med. 37(2) (2007), pp. 251–261.
[53] M. Shirahata, K.I. Koizumi, S. Saito, N. Ueno, M. Oda, N. Hashimoto, and K. Kato, *Gene expression based molecular diagnostic system for malignant gliomas is superior to histological diagnosis*, Clin. Cancer Res. 13 (2007), pp. 7341–7356.

[54] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub and W.R. Sellers, *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell 1(2) (2002), pp. 203–209.
[55] A.I. Su, J.B. Welsh, L.M. Sapinoso, S.G. Kern, P. Dimitrov, H. Lapp, P.G. Schultz, S.M. Powell, C.A. Moskaluk, H.F. Frierson and G.M. Hampton, *Molecular classification of human carcinomas by use of gene expression signatures*, Cancer Res. 61(20) (2001), pp. 7388–7393.
[56] R. Tibsrani, T. Hastie, B. Narasimhan, and G. Chu, *Diagnosis of multiple cancer types by shrunken centriods of gene expression*, Proc. Natl. Acad. Sci. USA 99 (2002), pp. 6567–6572.
[57] L.J. Van't Veer, H. Dai, M.J. vande Vijver, Y.D. He, A.A. Hart, M. Mao, S.H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer*, Nature 415 (2002), pp. 530–536.
[58] L Wang, F. Chu, and W. Xie, *Accurate cancer classification using expression of very few genes*, IEEE/ACM Trans. Comput. Biol. Bioinformatics 4(1) (2007), pp. 40–53.
[59] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics 1 (1945), pp. 80–83.
[60] K. Yang, Z. Cai, J. Li, and G. Lin, *A stable gene selection in microarray data analysis*, BMC Bioinformatics 7 (2006), pp. 228.
[61] H.L. Yu, S. Gao, B. Qin, and J. Zhao, *Multiclass microarray data classification based on confidence evaluation*, Genet. Mol. Res. 11(2) (2012), pp. 1357–1369.
[62] Y. Zhang, C. Ding, and T. Li, *A two-stage gene selection algorithm by combining reliefF and mRMR*, BMC Genom. 9 (Suppl 2) s27, pp. 164–171.
[63] Z. Zhu, Y.S. Ong, and M. Dash, *Wrapper–filter feature selection algorithm using a memetic framework*, IEEE Trans. Syst. Man Cybern.—Part B: Cybernetics 37(1), 2007, pp. 70–76.