Liu et al. BMC Bioinformatics (2018) 19:396
https://doi.org/10.1186/s12859-018-2400-2

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Feature selection of gene expression data for Cancer classification using double RBF-kernels

Shenghui Liu[1], Chunrui Xu[1,2], Yusen Zhang[1*], Jiaguo Liu[1], Bin Yu[3], Xiaoping Liu[1*] and Matthias Dehmer[4,5,6]

## Abstract

**Background:** Using knowledge-based interpretation to analyze omics data can not only obtain essential information regarding various biological processes, but also reflect the current physiological status of cells and tissue. The major challenge to analyze gene expression data, with a large number of genes and small samples, is to extract disease-related information from a massive amount of redundant data and noise. Gene selection, eliminating redundant and irrelevant genes, has been a key step to address this problem.

**Results:** The modified method was tested on four benchmark datasets with either two-class phenotypes or multiclass phenotypes, outperforming previous methods, with relatively higher accuracy, true positive rate, false positive rate and reduced runtime.

**Conclusions:** This paper proposes an effective feature selection method, combining double RBF-kernels with weighted analysis, to extract feature genes from gene expression data, by exploring its nonlinear mapping ability.

**Keywords:** Clustering, Gene expression, Cancer classification, Feature selection, Data mining

## Background

Gene expression data can reflect gene activities and physiological status in a biological system at the transcriptome level. Gene expression data typically includes small samples but with high dimensions and noise [1]. A single gene chip or next generation sequencing technology can detect at least tens of thousands of genes for one sample, but when it comes to some diseases or biological processes, only a few groups of genes are related [2, 3]. Moreover, testing these redundant genes not only demands tremendous search space but also reduces the performance of data mining due to the overfitting problem. Thus, extracting the disease-mediated genes from the original gene expression data has been a major problem for medicine. Moreover, the identification of appropriate disease-related genes will allow the design of relevant therapeutic treatments [4, 5].

So far, several feature selection methods have been suggested to extract disease-mediated genes [6–8]. Zhou et al. [3] proposed a new measure, LS bound measure, to address numerous redundant genes. Several statistical theories ($\chi^2$et al.) and classic classifiers (Support Vector Machine et al.) have been used in feature selection [9]. In general, these methods can be divided into three categories: filter, wrapper and embedded methods [9, 10]. The filter method is based on the structural information of the dataset itself, which is independent of the classifier, and it selects a feature subset from the original dataset using a certain evaluation rule based on statistical methods [11]. The wrapper method [12] is based on the performance of the classifier to evaluate the significance of feature subsets, while the embedded method [13] combines the advantage of filter and wrapper methods, selecting feature genes using a pre-determined classification algorithm [14, 15]. Since the filter methods are independent of the classifier, the computational complexity of these methods is relatively low, hence, they are suitable for massive data processing [16]. Yet, wrapper

* Correspondence: zhangys@sdu.edu.cn; xpliu@sdu.edu.cn
[1]School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China
Full list of author information is available at the end of the article

methods can reach a higher accuracy, but they also have a higher risk of over-fitting.

Kernel methods have been one of the central methods in machine learning in recent years. They have widely been applied to the area of classification and regression. A kernel method has the capability of mapping the data (non-linearly) to a higher dimensional space [17]. Hence, by using the kernel method, the dimension of the observed data such as gene expression data can be significantly reduced, that is, the irrelevant genes can be filtered by kernel method, thus revealing the hidden inherent law in the biological system [18]. Characteristically, kernels have a great impact on learning and predictive results of machine learning methods [5, 19].

Although a great number of kernels exist and it is intricate to explain their distinctive characteristics, kernels used by feature extraction can be divided into two classes: global and local kernels, such as polynomial and radial basis function (RBF) kernels. The influence of different types of kernels on the interpolation and extrapolation capabilities has been investigated. In global kernels, data points far away from the test point have a profound effect on kernel values, while, by using local kernels, only those close to the test point have a great effect on kernel values. The polynomial kernel shows better extrapolation abilities at lower orders of the degrees, but requires higher orders of degrees for good interpolation, while the RBF-kernel has good interpolation abilities, but fails to provide longer range extrapolation [17, 20].

KBCGS [20] is a new filter method based on the RBF-kernel using weighted gene measures in clustering. This supervised learning algorithm applied global adaptive distance to avoid falling in local minima. The RBF kernel function has been proven useful when it comes to show a satisfactory global classification performance for gene selection. Yet, exploring this problem in depth definitely needs further research. A typical mixture kernel is to construct a convex combination of basis kernels. Based on the characteristics of the original kernel function, linear fusion of a local kernel function and a global kernel function can constitutes a new mixed kernel function. Several mixture kernels have been introduced in [21–23] to overcome limitations of single-kernel, which can enhance the interpretability of the decision, function and improve performance. Phienthrakul et al. proposed Multi-scale RBF Kernels in Support Vector Machines and demonstrated that the use of Multi-scale RBF Kernels could result in better performance than that of a single RBF on benchmarks [23].

In this paper, we modified KBCGS based on double RBF-kernels, and applied the proposed method to feature selection of gene expression. We introduced the double RBF-kernel to both SVM and KNN, and evaluated their performance in the area of gene selection. This mixture describes varying degrees of local and global characteristics of kernels only by choosing different values of $\gamma_1$and $\gamma_2$. We combined the double RBF-kernel with a weighted method to overcome the limitations of single and local kernel. As an application, we provided a feature extraction method which uses this kernel, applying our method to several benchmark datasets: diffuse large B-cell lymphoma (DCBL) [24], colon [2], lymphoma [1], gastric cancer [25], and mixed tumors [26] to evaluate its performance. The results demonstrate that this method allows better discrimination in gene selection. In addition, the method is superior when it comes to accuracy and efficiency if we compare this technique with traditional gene selection methods.

This paper provides a brief overview of the gene selection method for expression data analysis, then, the improved KBCGS method called DKBCGS (Double-kernel KBCGS), in which the two classification methods were used for the clustering analysis was compared to six popular gene selection methods. The last section of the paper provides a comprehensive evaluation of the proposed method using four benchmark gene expression datasets.

## Methods

Gene expression data with $l$ genes and $n$ samples can be represented by the following matrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1l} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nl} \end{bmatrix} \quad (1)$$

$X_i$is a row vector that represents the total gene expression levels of sample $i$ and $x_{ij}$ is the expression level of gene $j$ of sample $i$.

### Cluster center

In this paper, we used Z-score to normalize the original data. The standard score $Z$ used for a gene is as follows:

$$Z = \frac{(x-\mu)}{\sigma} \quad (2)$$

where, $x$ is the expression level of a gene in a sample, $\mu$ is the mean value of the gene across all samples, and $\sigma$ is its standard deviation of the gene across all samples.

The cancer classification was formulated as a supervised learning problem, defining the cluster center as:

$$v_{ik} = \frac{1}{|C_i|} \sum_{X_j \in C_i} x_{jk} \tag{3}$$

In this equation, $I = 1, 2,..., C, j = 1,2,...,n, k = 1,2,...,l$, $C_i$ is the number of samples contained in class $C_i$, respectively. Hence, $V_i = [v_{i1},...,v_{il}]$ is the cluster center of class $C_i$.

### Double RBF-kernels

The kernel function acts as a similarity measure between samples in a feature space. A simple form of similarity measure is the dot product between two samples. The most frequently used kernel is a positive definite Gaussian kernel [27]. The classic Gaussian kernel on two samples $x$ and $x_i$, represented as feature vectors in an input space, is defined by:

$$K_{rbf}(x, x_i) = e^{-\gamma_1 \|x - x_i\|^2} \tag{4}$$

where, $\gamma_1 > 0$ is a free parameter.

It is a positive definite kernel representing local features, therefore, it can also be used as the kernel function to weight genes for the gene selection method. Kernel methods have already been applied to many areas due to their effectiveness in feature selection and dimensionality reduction [27]. However, for the purposes of these methods, the focus is on creating a more general unified mixture kernel that has capabilities of both local and global kernels.

This work utilizes a double RBF-kernel as a similarity measure. The number choice of kernels could typically depend on the level of heterogeneity of the datasets. Increasing numbers of kernels helps to improve accuracy, but increase the computational cost. Therefore, we have to find a compromise between multiple kernels learning and double RBF-kernel learning, based on the performance and computational complexity. In most case, two RBF kernels are enough to handle most data with reasonable accuracy and computational cost. It should be emphasized that the proposed nonlinear kernel method is based on the combination of two RBF-kernels that has few limitations when calculating the distance among genes as follows:

$$K_{\gamma_1\gamma_2}(x, x_j) = ce^{-\gamma_1 \|x - x_i\|^2} + (1-c)e^{-\gamma_2 \|x - x_i\|^2} \quad (\gamma_1 > 0, \gamma_2 > 0) \tag{5}$$

To further illustrate Eq. (5), the mapping relationships were plotted between the formula Eq. 5 and RBF-kernel by Figs. 1 and 2. Figures 1 and 2 clearly show the fat-tailed shape of the mapping changes with $\gamma_1$ , $\gamma_2$ and compared to the RBF mapping parameter $\gamma_1$. Figure 2 shows changing parameters $\gamma_1$ , $\gamma_2$, the lower graph varies more slightly than the upper one. Therefore, the double-kernel can fit data better with less impact by outliers, indicating that the double-kernel has better flexibility than the single-kernel. The fat- tail characteristics make the double RBF kernels have better learning ability and better generalization ability than a RBF-kernel.

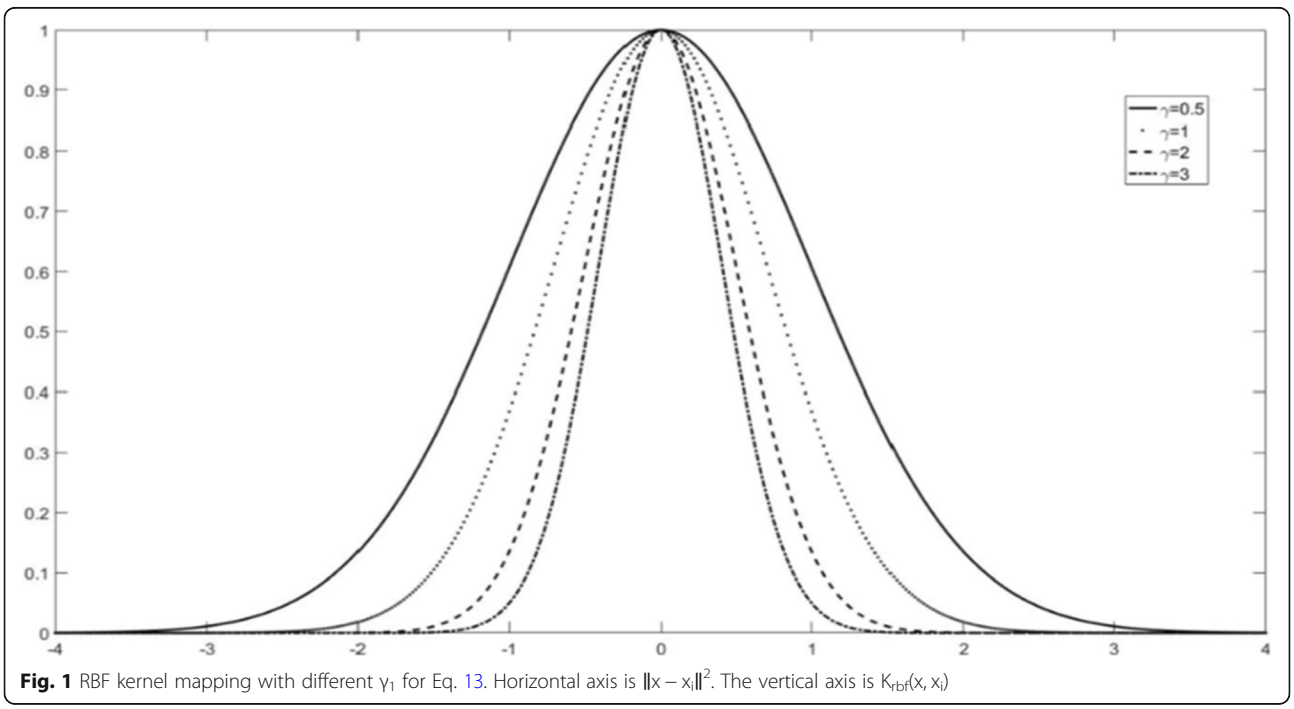


**Fig. 1** RBF kernel mapping with different $\gamma_1$ for Eq. 13. Horizontal axis is $\|x - x_i\|^2$. The vertical axis is $K_{rbf}(x, x_i)$

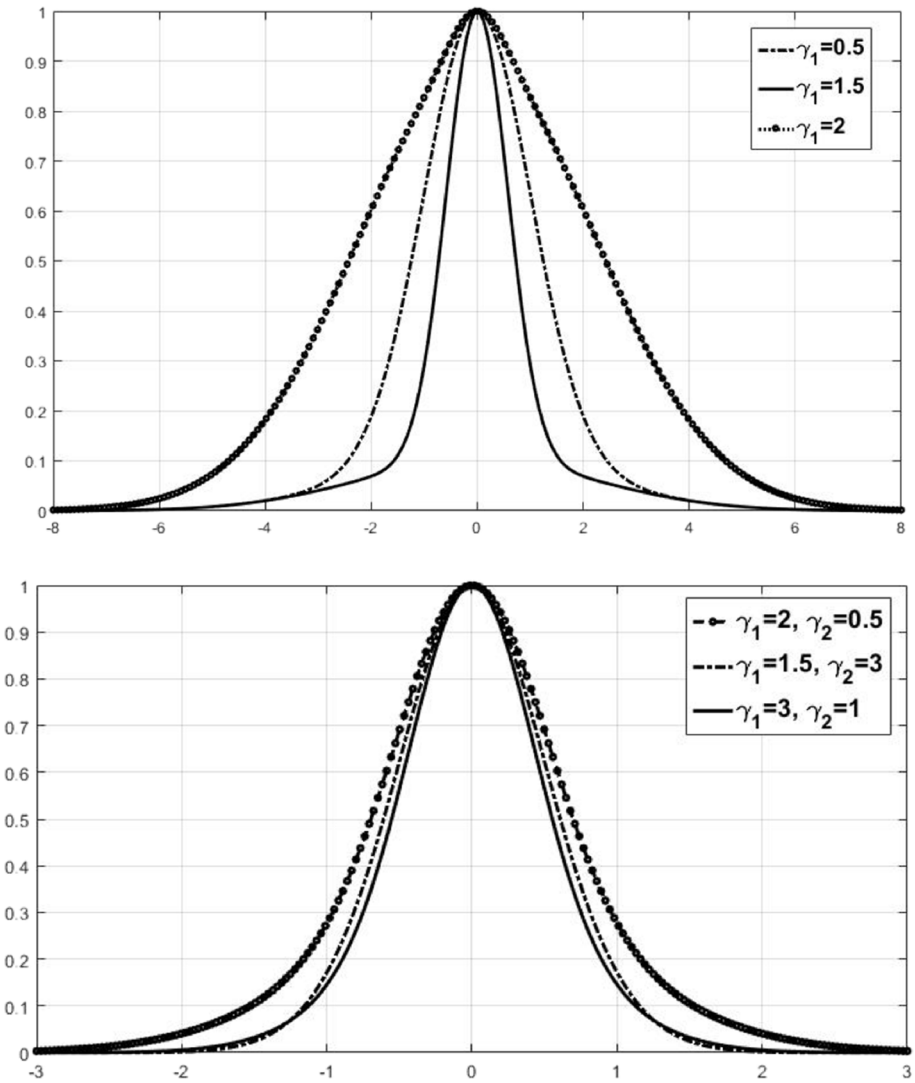


**Fig. 2** The mapping with different $\gamma_1$ and $\gamma_2$ for Eq. (5). The first figure is for $\gamma_1$ only, and the second figure is for the combination of $\gamma_2$ and $\gamma_1$. The horizontal axis is given by $\|x - x_i\|^2$ and the vertical axis is given by $K_{\gamma_1\gamma_2}(x, x_j)$

### Kernels as measures of similarity

Suppose $\Phi : X \longrightarrow F$ is a nonlinear mapping from the space X to a higher dimensional space F, By applying the mapping Φ, then the dot product $x_k^T x_l$ in the input space X is mapped to $\Phi(x_k)^T \Phi(x_l)$ in the new feature space. The key idea in kernel algorithms is that the non-linear mapping Φ doesn't need to be explicitly specified because each Mercer kernel can be expressed as:

$$K(x_k, x_l) = \Phi(x_k)^T \Phi(x_l) \tag{6}$$

that is usually referred to as kernel trick [22]. Then, the Euclidean distances in F yields:

$$\begin{aligned} \|\Phi(x_k) - \Phi(x_l)\|^2 &= (\Phi(x_k) - \Phi(x_l))^T (\Phi(x_k) - \Phi(x_l)) \\ &= K(x_k, x_k) - 2K(x_k, x_l) + K(x_l, x_l) \end{aligned} \tag{7}$$

Then, a dissimilarity function between an sample and a cluster centroid could be defined as:

$$\begin{aligned} \phi^2(x_j, v_i) &= \sum_{k=1}^{l} \left\| \Phi(x_{jk}) - \Phi(v_{ik}) \right\|^2 \\ &= \sum_{k=1}^{l} \left( K(x_{jk}, x_{jk}) - 2K(x_{jk}, v_{ik}) + K(v_{ik}, v_{ik}) \right) \end{aligned} \tag{8}$$

### Gene ranking and selection

The most used gene selection methods belong to the so-called filter approach. Filter-based feature ranking

methods rank genes independently without any learning algorithm. Feature ranking consists of weighting each feature according to a particular method, then selecting genes based on their weights.

In this paper, our method DKBCGS is based on a KBCGS method improved to achieve higher accuracy and converge faster.

The KBCGS method adopted global distance, assigning different weights to different genes. The clustering objective function is given by:

$$\begin{aligned} J &= \sum_{i=1}^{C}\sum_{x_j\in C_i}\phi^2(X_j, V_i) + \delta\sum_{k=1}^{l} W_k^2 \\ &= \sum_{i=1}^{C}\sum_{x_j\in C_i}\sum_{k=1}^{l} W_k\left\|\Phi(X_{jk})-\Phi(V_{ik})\right\|^2 \\ &+\delta\sum_{k=1}^{l} W_k^2 \end{aligned} \tag{9}$$

where $w = (w_1, w_2,...,w_l)$ are the weight of genes.

$$\begin{cases} w_k\in[0,1], k = 1,2,\ldots,1 \\ \sum_{k=1}^{l} w_k = 1 \end{cases} \tag{10}$$

As shown in Eq. (1), the first part is the sum of weighted dissimilarity distance among samples and the cluster they belong to evaluated by the kernel method. This part will reach its minimum value only when there is one gene that is completely relevant and the others are irrelevant. The second part is the sum of squared weights of genes, which will only reach its minimum value when all genes are equally weighted. Therefore, by combining these two parts, the optimal gene weights are obtained, then the feature genes can be selected.

To minimize J with respect to the restriction Eq. (10), the Lagrange multipliers methods were applied as follows:

$$J(w_k,\lambda) = \sum_{i=1}^{C}\sum_{x_j\in C_i}\phi^2(x_j, v_i) + \delta\sum_{k=1}^{l} w_k^2-\lambda\left(\sum_{k=1}^{l} w_k-1\right) \tag{11}$$

So, the partial derivative of $J(w_k,\lambda)$ is given by:

$$\begin{cases} \frac{\partial J(w_k,\lambda)}{\partial\lambda} = \sum_{k=1}^{l} w_k-1 \\ \frac{\partial J(w_k,\lambda)}{\partial w_k} = \sum_{i=1}^{C}\sum_{x_j\in C_i}\left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2 + 2\delta w_k-\lambda \end{cases} \tag{12}$$

The $J(w_k,\lambda)$ reaches its minimum when the value of the partial derivative is zero. So, w is calculated as follows:

$$\begin{aligned} w_k = &\frac{1}{l} + \frac{1}{2\delta}\sum_{i=1}^{C}\sum_{x_j\in C_i} \\ &\left(\frac{\sum_{i=1}^{C}\sum_{x_j\in C_i}\left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2}{1} - \left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2\right) \end{aligned} \tag{13}$$

Based on Eq. (13), the KBCGS method chooses $\frac{1}{l}$ as the initial weight of $w_k$. In the second part of Eq. (9), the choice of δ is quite important since it represents the distance of genes. The value of δ should ensure that both parts are of the same order of magnitude, so according to SCAD algorithm [28], the δ is calculated iteratively as follows:

$$\delta^{(t)} = \alpha\frac{\sum_{i=1}^{C}\sum_{x_j\in C_i}\sum_{k=1}^{l} w_k^{(t-1)}\left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2}{\sum_{k=1}^{l}\left(w_k^{(t-1)}\right)^2} \tag{14}$$

Where α is a constant which influences the value of δ, with a default value of 0.0S5. The Gaussian kernel is employed in this algorithm:

$$K_{rbf}(x, x_i) = e^{-\gamma_1\|x-x_i\|^2} \tag{15}$$

Where, $\gamma_1 > 0$ is a free parameter and the distance can be expressed as:

$$\left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2 = 2(1-K(x_{jk}, v_{ik}) \tag{16}$$

The max number of iteration is 100, and $\theta = 10^{-6}$.

The features of the improved method are outlined below. Similar to KBCGS algorithm [20], the clustering objective function is defined:

$J = \sum_{i=1}^{C}\sum_{x_j\in C_i}\phi^2(x_j, v_i) + \delta\sum_{k=1}^{l} w_k^2$

where $w = (w_1, w_2,...,w_l)$ are the weight of genes.

The DKBCGS method calculates δ iteratively according to Chen's approach [20], however, it is improved the iterative method to calculate w by deriving the following formula:

$$\delta^{(t)} = \left|\frac{\sum_{i=1}^{C}\sum_{xj\in C_i}\sum_{k=1}^{l} w_k^{(t-1)}\left\|\Phi(x_{jk})-\Phi(v_{ik})\right\|^2}{\sum_{k=1}^{1}\left(w_k^{(t-1)}\right)^2}\right| \tag{17}$$

and instead of Gaussian kernel, the double RBF-kernel is used as mentioned in Eq. (5).

The initial value of δ in Eq. (13) is important in our algorithm since it reflects the importance of the second term relative to the first term. If δ is too small, the only one feature in cluster i will be relevant and assigned a weight of one. All other feature will be assigned zero weights. On the other hand, if δ is too large, then all feature in cluster I will be relevant, and assigned equal weights of 1/n. The values

of δ should be chosen such that both terms are of same order of magnitude. In all examples described in this paper, we compute δ iteratively using Eq. (17) as SCAD method, see [28].

Through improving the iteration method, we achieve less iteration, therefore an improvement toward convergence compared to the KBCGS method. As previously mentioned, gene expression datasets are often linearly non-separable, so choosing an appropriate nonlinear kernel to map the data to a higher dimensional space has been proven efficient.

### Implementation

The algorithm can be stated using the following pseudocode:

Input: Gene expression dataset X and class label vector y;

Output: weights vector w of genes;

Use Z-score to normalize the original data X;

Use Eq. (3) to calculate the cluster center of different class of genes in the input space, respectively;

Use Eq. (8) to calculate the dissimilarity between the genes and their cluster center of class;

Initial value: $w_0 = \frac{1}{I}$;

Repeat:

Use Eq. (14) to find the $(t+1)$th distance parameter $\delta^{(t+1)}$;

Use Eq. (13) to calculate $(t+1)$th weights $w^{(t+1)}$ of genes;

Use Eq. (11) to calculate $(t+1)$th objective function $J^{(t+1)}$;

Until: $J^{(t+1)} - J^{(t)} < \theta$.

Return $w^{(t+1)}$.

We constructed SVM and KNN classifiers for each dataset. These methods have been introduced in the Additional file 2. A 10-fold cross validation was used as the validation strategy to reduce the error and obtain classification accuracy.

The whole experiment was performed using MATLAB. To determine the value of hyperparameters, we use the grid search method. Figure 3 shows the change of in the average error rate with the change in the number of selected feature genes by employing DKBCGS. It is obvious that there is a great improvement in the results when the selected feature genes number increases from 1 to 20. In order to identify the optimal performance of all datasets, the number was restricted from 1 to 50.

## Results

To validate the performance of DKBCGS method, it was compared with some commonly used filter-based feature ranking methods namely $\chi^2$-Statistic, Maximum relevance and minimum redundancy (MRMR), Relief-F, Information Gain and Fisher Score. These methods have been introduced in the Additional file 1. Also, the improved approach was compared with KBCGS [20].

### Dataset description

The four datasets used as benchmark examples in this work are shown in Table 1. The specifics of these datasets are outlined in the Additional file 3.

## Discussion

By using the two-class datasets, the performance of proposed method, in comparison to the other six methods, was evaluated by calculating the accuracy (ACC), the true positive rate (TPR) and the true negative rate (TNR).

Table 2 and Table S1 shows the results of the two-class datasets. These results indicate that the proposed method has high accuracy and short runtime in both the SVM and KNN classifier, while MRMR also performs well in the KNN classifier. Fig. S1 tell us that the expression of the characteristic genes selected by the proposed algorithm has significant differences in the expression level of normal/diseased samples.

Gene-set enrichment analysis is used to identify coherent gene-sets. Fig. 5 show us that the genes (dataset: colon

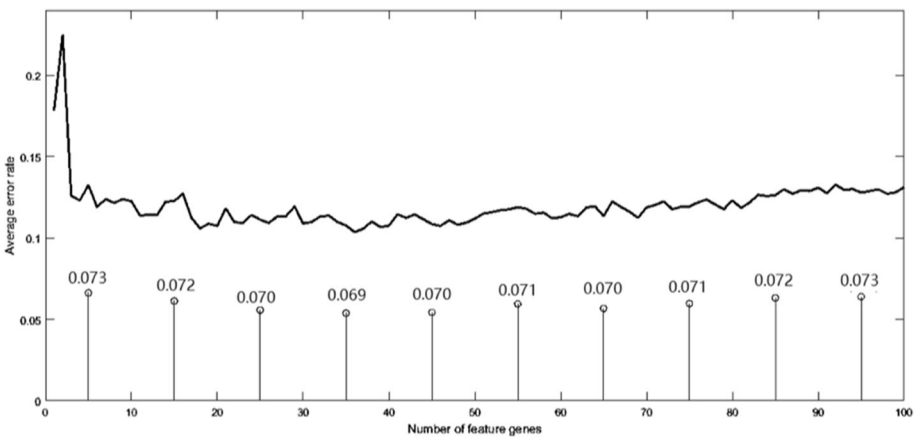


**Fig. 3** Average error rate versus different number of selected feature genes

**Table 1** Summary of the four gene expression datasets

| | Samples | Classes | Genes | References |
|---|---|---|---|---|
| DLBCL | 77 | 2 | 7129 | Shipp et al. [24] |
| Gastric cancer | 40 | 2 | 1519 | Boussioutas et al. [25] |
| Multi-cancer | 152 | 5 | 65,522 | Yuan et al. [26] |
| Lymphoma | 62 | 3 | 4026 | Alizadeh et al. [1] |

cancer), selected by DKBCGS, enriched in strongly connected gene-gene interaction networks and in highly significant biological processes. Furthermore, the significant difference between the expression profiles for the top-ranked genes selected by DKBCGS in the form of a color map in Fig. 6 (a) and the expression profiles for eight genes chosen randomly from the base is presented in Fig. 6 (b) confirms the good performance of the proposed selection procedure.

### Classification accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad 0 \leqslant ACC \leqslant 1 \qquad (18)$$

TP, TN, FP, FN are the True Negatives, True Positives, False Negatives and False Positives, respectively.

As the number of positive samples and negative samples using the two-class datasets are not equal, the true positive rate (TPR) and the true negative rate (TNR) were used as another strategy for measuring the performance, considering both the precision and the recall of the experiment under test. Precision represents the number of correct positive results divided by the number of all positive results. Recall is the number of correct positive results divided by the number of positive results that should have been returned. Therefore, the TPR and false positive rate (FPR) are calculated as follows:

### True positive rate

$$\mathrm{TPR} = \frac{TP}{TP + FN} \qquad (19)$$

### True negative rate

$$\mathrm{TNR} = \frac{TN}{FP + TN} \qquad (20)$$

**Table 2** Performance of gene feature selection methods with KNN classifier (high) and SVM classifier (low) in two-class datasets

| Dataset: Gastric cancer | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | 0.9821 | 0.9664 | 0.9875 | 0.9779 | 0.9038 | 0.9548 | **0.9986** | 0.9716 |
| TNR | 1.0000 | 0.9500 | 0.9367 | 1.0000 | 0.9500 | 0.9800 | 1.0000 | 0.9755 |
| TPR | 0.9818 | 0.9677 | 0.9969 | 0.9759 | 0.8771 | 0.9498 | 1.0000 | 0.9826 |
| TIME(s) | **0.0846** | 0.7349 | 1.4736 | 0.7542 | 9.7452 | 4.2604 | 0.9007 | 0.6518 |
| Dataset: DLBCL | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | **0.9833** | 0.9615 | 0.9865 | 0.9712 | 0.9123 | 0.9245 | 0.9341 | 0.9795 |
| TNR | 0.9943 | 0.9456 | 0.9422 | 0.9854 | 0.9457 | 0.9456 | 0.9654 | 1.0000 |
| TPR | 0.9863 | 0.9513 | 0.9645 | 0.9541 | 0.9024 | 0.9234 | 0.9432 | 0.9712 |
| TIME(s) | 0.1215 | 0.2257 | 0.1954 | 0.1857 | 0.1678 | 0.5111 | **0.0931** | 0.2148 |
| Dataset: Gastric cancer | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | **1.0000** | 0.9768 | 0.9855 | 0.9623 | 0.9168 | 0.973 | 0.9988 | 0.9822 |
| TNR | 1.0000 | 0.9611 | 0.95 | 0.9158 | 0.9316 | 0.9433 | 1.0000 | 1.0000 |
| TPR | 1.0000 | 0.9929 | 0.9971 | 0.9776 | 0.9121 | 0.9827 | 1.0000 | 0.9755 |
| TIME(s) | **0.0846** | 0.7349 | 1.4736 | 0.7542 | 9.7452 | 4.2604 | 0.9007 | 0.7418 |
| Dataset: DLBCL | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | **1.0000** | **1** | 0.9975 | 1.0000 | 0.9975 | 0.9750 | 0.9975 | 0.9845 |
| TNR | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9683 | 0.9733 | 0.9571 | 0.9579 |
| TPR | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.8383 | 0.9437 | 0.9917 | 0.9931 |
| TIME(s) | 0.1215 | 0.2257 | 0.1954 | 0.1857 | 1.6478 | 0.5111 | **0.0931** | 0.2148 |

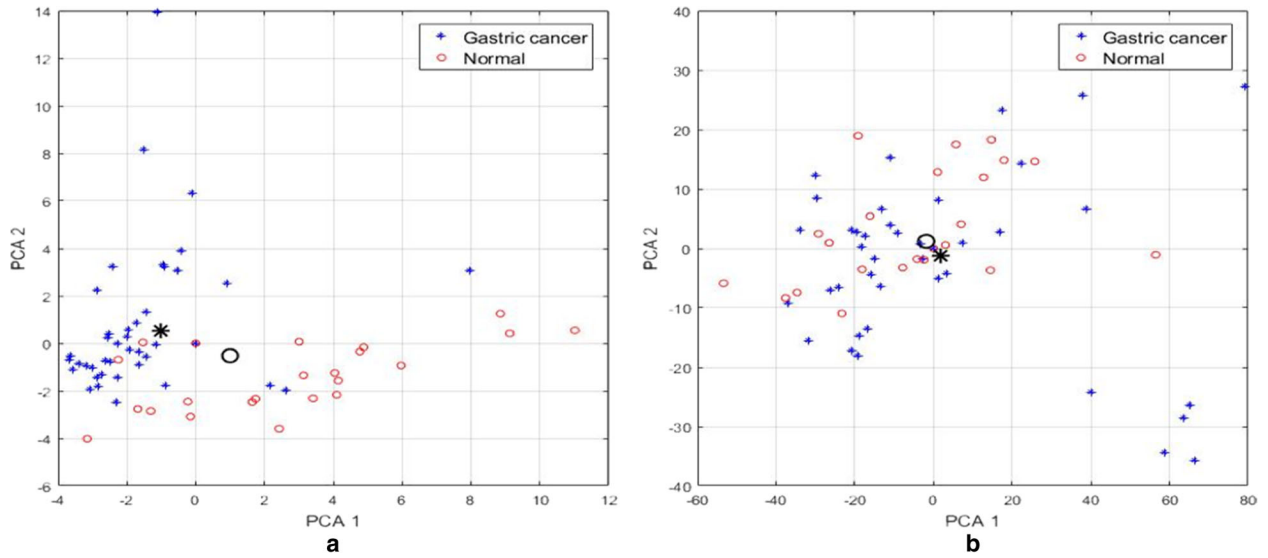


**Fig. 4** The distribution of the two-class samples mapped on the two most important principal components at representation of vectors x by 50 most significant genes (**a**) and at application of all genes (**b**). The horizontal axis is the first principal component and the vertical axis is the second principal component. Black marks represent different categories of the centers

Table 2 shows the results of the two-class datasets. The runtime of DKBCGS, being less than 0.1 s, is much shorter than others, except for runtime of MRMR-SVM in the DLBCL dataset, that is, the proposed double-kernel model can efficiently reduce computation complexity. Regarding accuracy, the proposed method also performs well, reaching 100% in SVM classifier and slightly less than that of MRMR in KNN classifier. Taken together, these results indicate that the proposed method has high accuracy and short runtime in both the SVM and KNN classifier, while MRMR also performs well in the KNN classifier. Also, the average ROC (Receiver Operating Characteristic) curve was plotted for further evaluation in Fig. 4. A further comparison with KBCGS in four datasets, calculating average results of KNN and SVM, is shown in Additional file 4: Table S1. The results clearly demonstrate that the improved approach DKBCGS performs better in both runtime and accuracy.

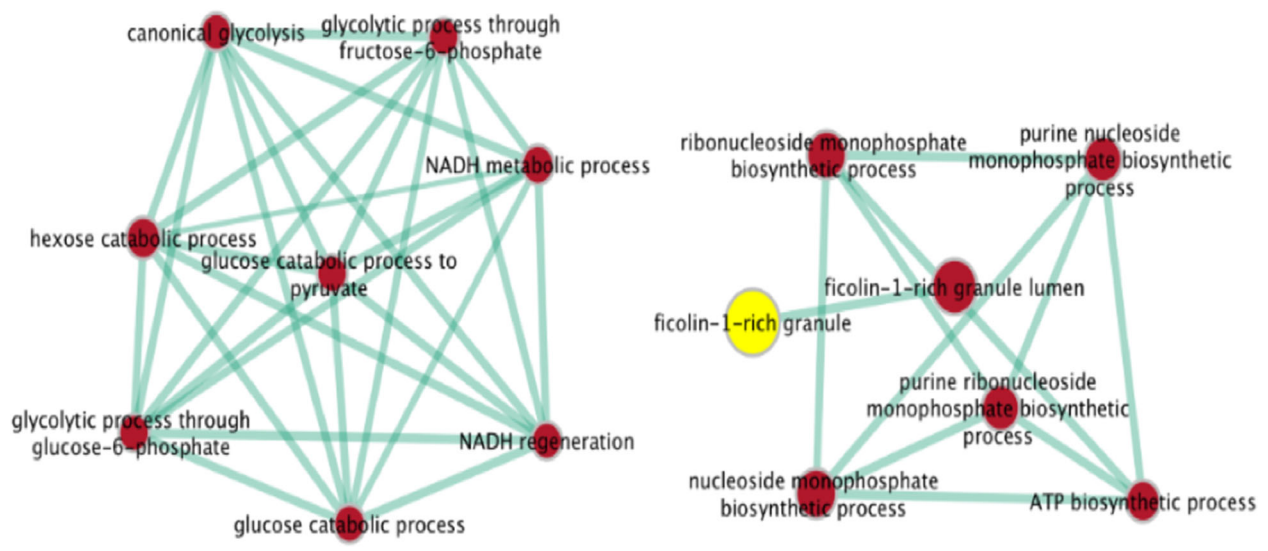


**Fig. 5** GO Enrichment Mapping the cluster-specific genes for the DLBCL dataset (P-value < 0.001). We firstly identified significant GO terms on the g: profiler web interface. Then we used the enrichment map plug-in in Cytoscape [29] to visualize these significant GO terms. Each node represents a GO term and each edge represents the degree of gene overlap (Jaccard similarity) that exists between two gene sets corresponding to the two GO terms

Regarding the gastric cancer dataset, we have mapped the multidimensional observations into 2-dimensional space formed by the two most important principal components.

Two cases have been investigated. The first approach deals with using the original vectors only containing 50 genes selected by the fusion procedure. Fig. 5(a) depicts this case in which only the best representative genes in the vector x are used. For comparison, the Principal component analysis (PCA) was repeated for the full-size original 2000 element vectors containing all genes. The graphical results of the sample distribution are presented in Fig. 5(b). Large bold symbols of the circle and x represent the centroids of the data belong to two classes.

Furthermore, the first fifty top-ranked gene expression levels were analyzed in the gastric cancer dataset using the various methods as shown in Additional file 5: Figure S1. It can be clearly seen that the expression of the characteristic genes selected by the proposed algorithm has significant differences in the expression level of normal/diseased samples, therefore has some research value.

### Gene-set enrichment analysis

Gene-set enrichment analysis is useful to identify coherent gene-sets, such as pathways, that are statistically overrepresented in a given gene list. Ideally, the number of resulting sets is smaller than the number of genes in the list, thus simplifying interpretation. However, the increasing number and redundancy of gene-sets used by many current enrichment analysis resources work against this ideal. Gene-sets are organized in a network, where each set is a node and links the representative gene overlap between sets [26]. So, as to dataset DLBCL, the genes selected by DKBCGS enriched in strongly connected gene-gene interaction networks and in highly significant biological processes (Fig. 6).

To illustrate the results in a graphical form, the expression levels of the selected genes (dataset: colon cancer) are presented in Fig. 7(a). This figure shows the image of the expression profiles for the top-ranked genes selected by DKBCGS in the form of a colormap. The vertical axis represents observations and the horizontal axis represents the genes arranged according to their importance. There is a visible border between the cancer group and the normal group. For comparison purposes, the image of the expression profiles for eight genes chosen randomly from the base is presented in Fig. 7(b). There is a significant difference between both images, which confirms the good performance of the proposed selection procedure.

Both Table 3 and Table S2 show the results of the multiclass datasets. Both tables clearly show that the KBCGS can reduce runtime with high accuracy in other multiclass datasets. When using the lung cancer gene expression data, there is a substantial improvement in the accuracy of the classification using the double RBF-kernel algorithm for each of the feature subsets, which demonstrates that the KBCGS method can select the appropriate genes efficiently compared to other methods. For lung cancers, the feature genes selected by the double RBF-kernel algorithm also result in a higher accuracy. It not only improves the accuracy of the classification of gene expression data, but also identifies informative genes that are responsible for causing diseases. Therefore, the double RBF-kernel method is better than the X2-Statistics, MRMR, Relief-F, Information Gain, and Kruskal-Wallis test. Also, the significant difference between the

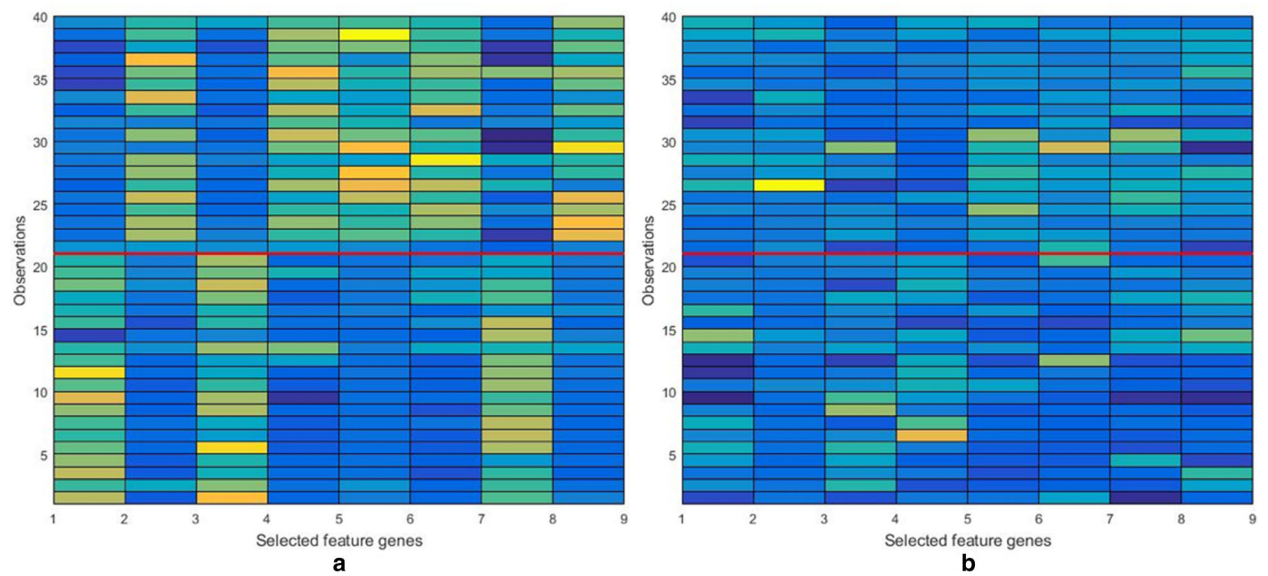


**Fig. 6** The colormap of the expression profiles for nine most significant genes selected by DKBCGS (**a**) and for 9 randomly chosen genes (**b**). The red line distinguishes between cancer samples and normal samples

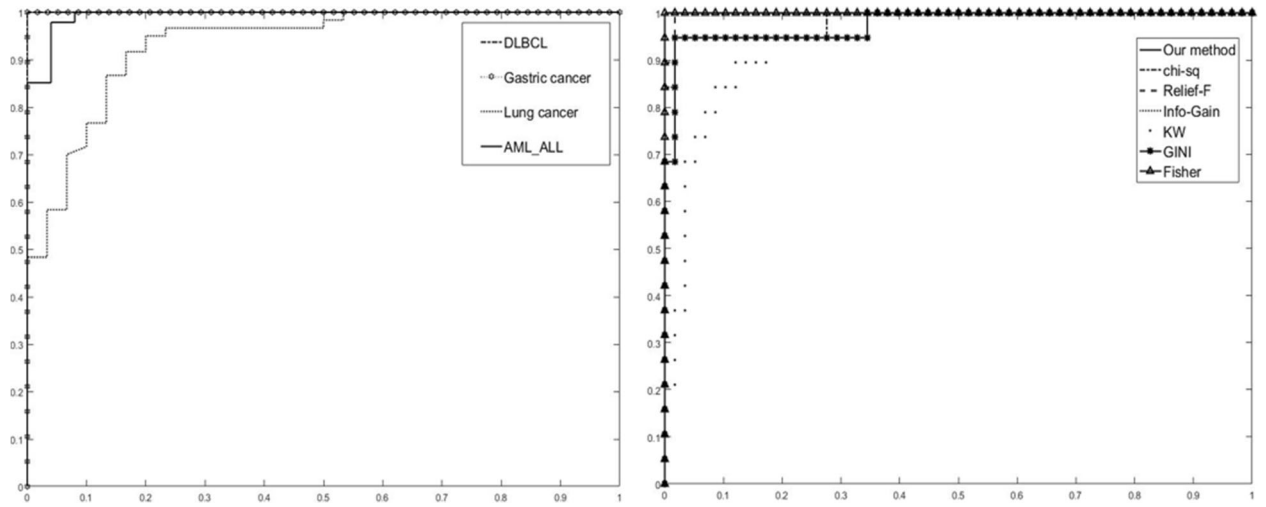


**Fig. 7** The ROC curve of two-class datasets, (left) ROC curve in different datasets and (right) shows the performance of different methods in DLBCL dataset. The horizontal axis is the false positive rate; the vertical axis is the true positive rate

**Table 3** Performance of gene feature selection methods with KNN classifier (high) and SVM classifier (low) in multiclass datasets

| Dataset: Lymphoma | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | **1.0000** | **1.0000** | **1.0000** | **1.0000** | 0.8617 | 0.9756 | **1.0000** | **1.0000** |
| Gene content | 25 | 70 | 26 | 49 | 22 | 16 | 29 | 35 |
| TPR | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.8617 | 0.9756 | 1.0000 | 1.0000 |
| TIME(s) | **0.3412** | 0.8944 | 2.3579 | 1.2561 | 7.4577 | 3.3144 | 1.5922 | 0.7541 |
| Dataset: Lung cancer | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | 0.9554 | 0.9443 | 0.9499 | **0.9641** | 0.9273 | 0.9472 | 0.9291 | 0.9514 |
| Gene content | 32 | 82 | 97 | 65 | 39 | 88 | 50 | 40 |
| TPR | 0.9243 | 0.9033 | 0.9185 | 0.9012 | 0.9210 | 0.9123 | 0.9042 | 0.9155 |
| TIME(s) | **0.1215** | 0.2250 | 0.1954 | 0.1857 | 1.6478 | 0.5111 | 0.2931 | 0.3171 |
| Dataset: Lymphoma | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | **1.0000** | 0.994 | 1.0000 | 1.0000 | 0.9283 | 0.9963 | 1.0000 | 1.0000 |
| Gene content | 35 | 34 | 34 | 16 | 28 | 17 | 27 | 40 |
| TPR | 1.0000 | 0.994 | 1.0000 | 1.0000 | 0.9283 | 0.9963 | 1.0000 | 1.0000 |
| TIME(s) | **0.3412** | 0.8944 | 2.3579 | 1.2561 | 7.4577 | 3.3144 | 1.5922 | 0.7541 |
| Dataset: Lung cancer | | | | | | | | |
| | DKBCGS | GINI | $X^2$-Statistic | Info.Gain | KW | RF | MRMR | KBCGS |
| ACC | 0.9151 | 0.9041 | 0.9115 | **0.9229** | 0.9102 | 0.9087 | 0.9199 | 0.9100 |
| Gene content | 64 | 87 | 75 | 89 | 71 | 60 | 77 | 74 |
| TPR | 0.9172 | 0.9005 | 0.9124 | 0.9285 | 0.9089 | 0.9114 | 0.9207 | 0.9122 |
| TIME(s) | **0.5736** | 1.8912 | 3.4551 | 2.4972 | 6.9322 | 4.1978 | 2.1207 | 1.0044 |

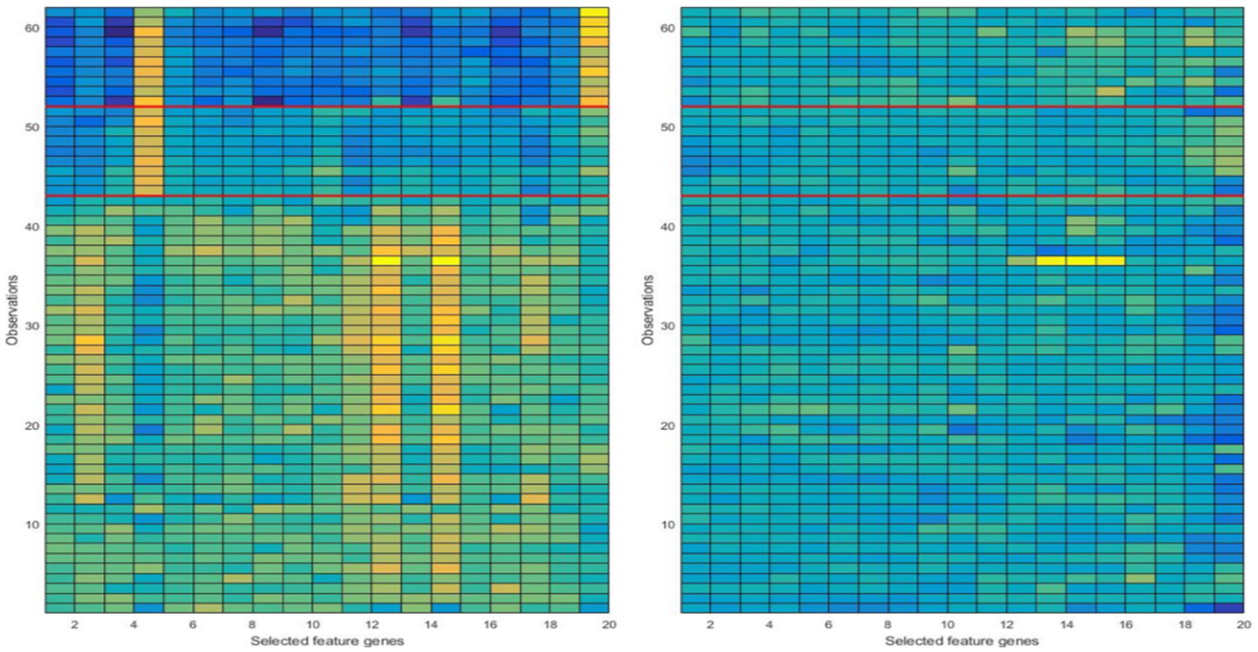


**Fig. 8** The colormap of the expression profiles for 20 most significant genes selected by the proposed method (left) and for 20 randomly chosen genes (right). The red line distinguishes between different classes

expression profiles for the top-ranked genes (dataset: Lymphoma) selected by DKBCGS in the form of a color map in Fig. 8 (a) and the expression profiles for 20 genes chosen randomly from the base is presented in Fig. 8 (b) demonstrates the good performance of the proposed selection procedure.

### Comparison of multiclass datasets

For the multiclass datasets, the performance of all methods was evaluated by computing accuracy (ACC) and run time (Time). The results are shown in Table 3. Also, further comparisons were made with KBCGS in other multiclass datasets, see Additional file 4: Table S2.

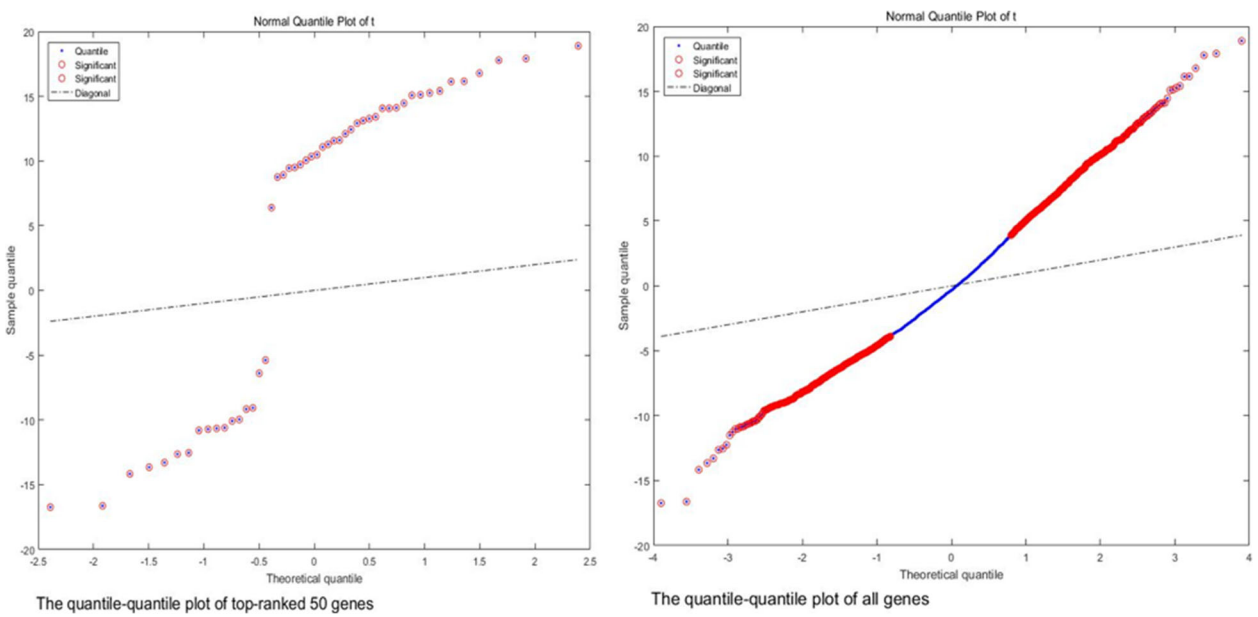


**Fig. 9** The quantile-quantile plot of top ranked 50 genes (left) and all genes (right). The horizontal axis is the theoretical quantile, the vertical axis is the sample quantile, and the red circle represents a significant gene

Both tables clearly show that the proposed method can reduce runtime with high accuracy.

When using the lung cancer gene expression data, there is a substantial improvement in the accuracy of the classification using the double RBF-kernel algorithm for each of the feature subsets, which demonstrates that the double RBF-kernel method can select the appropriate genes efficiently compared to other methods. For lung cancers, the feature genes selected by the double RBF-kernel algorithm also result in a higher accuracy. It not only improves the accuracy of the classification of gene expression data, but also identifies informative genes that are responsible for causing diseases. Therefore, the double RBF-kernel method is better than the $X^2$-Statistics, MRMR, Relief-F, Information Gain, and Kruskal-Wallis test. Also, the Information Gain method turns out to be highly competitive.

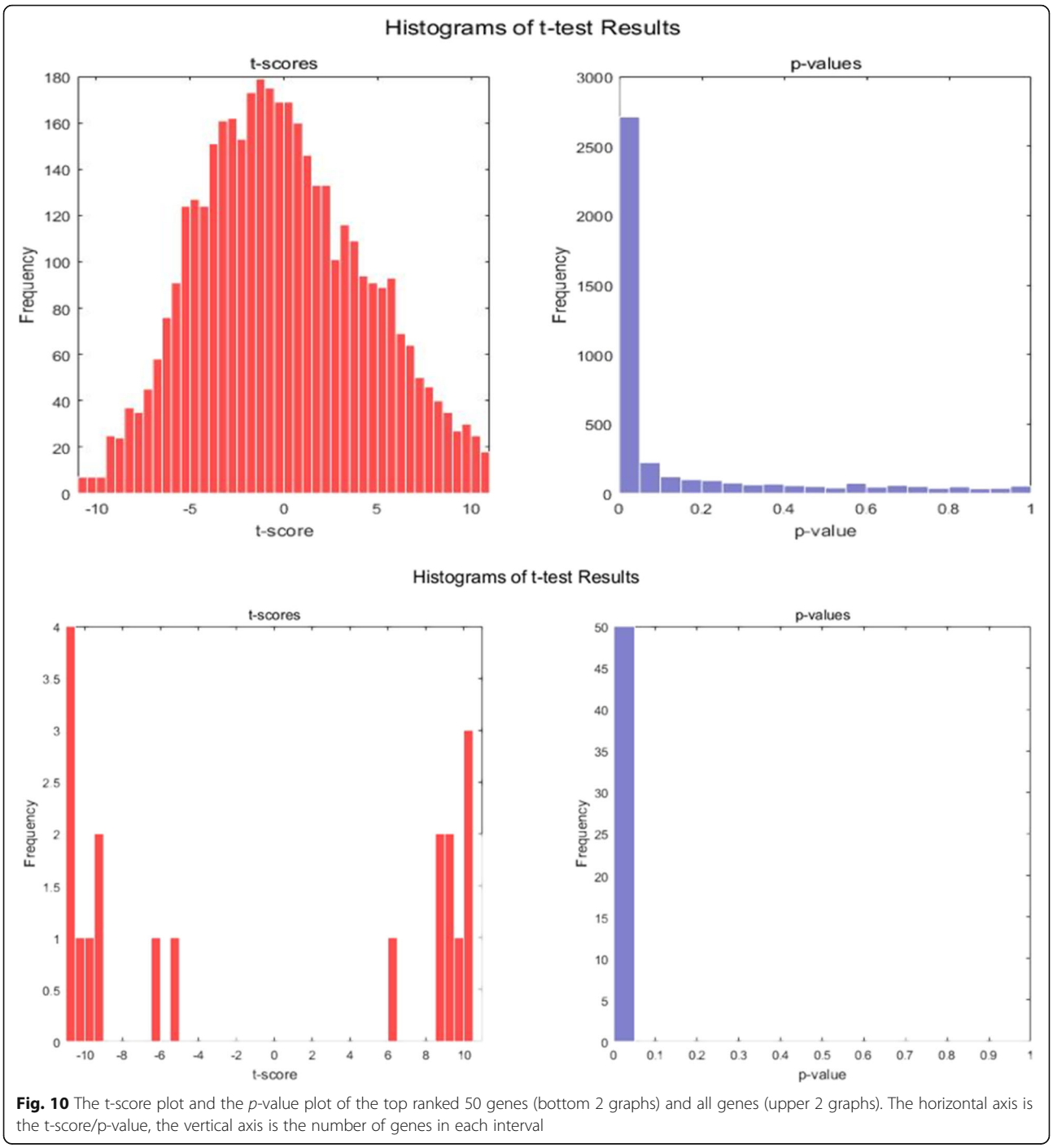


**Fig. 10** The t-score plot and the *p*-value plot of the top ranked 50 genes (bottom 2 graphs) and all genes (upper 2 graphs). The horizontal axis is the t-score/p-value, the vertical axis is the number of genes in each interval

In the second part of the experiment, the expression level of the selected genes (dataset: Lymphoma) was represented as before in Fig. 8(a). It shows the expression profiles for the top-ranked genes selected by fusion in the form of the colormap. There is a visible border between the different groups. Note that the images of the expression profiles for 20 genes are chosen randomly, see Fig. 8(b). There is a significant difference between both images, which demonstrates the performance of the proposed selection procedure.

### Differential gene expression analysis

The top 50 genes of Gastric cancer dataset were analyzed by applying the paired *t*-test method to obtain the t-score, *p*-value plot and the quantile-quantile plot of these genes. The quantile-quantile plot is mainly for identifying the gene expression levels of two classes. The results, as shown in Figs. 9 and 10, clearly show the difference between the feature genes obtained by DKBCGS and the original data. All the genes were divided into genes with significant attributes, and have a low *p*-value (average *p*-value = 0.023). Finally, this proves that DKBCGS has a certain statistical significance.

The t-score plot shows the normality of the data and the rationality of using the paired t-test. We can also conclude from the histogram of p-value that the paired t-test is significant because of the vast majority of p-value falls in the very end of the group of the histogram.

Between two groups of variables, a t-test is performed on each gene to identify significant differences in all genes and feature genes selected by our method, and a normal quantile map can be obtained by t-scores. A histogram of t-scores and *p*-values was used to study the test results.

## Conclusion

The number choice of kernels could typically depend on the level of heterogeneity of the datasets. Experiments on gene expression datasets show that double RBF-kernel outperforms all other used feature selection methods in terms of classification accuracies for both two-class datasets and multiclass datasets, especially in those datasets with small samples. The performances of double RBF-kernel learning in classification make it well suited alternatives to one RBF-kernel learning.

The use of known performance measures, such as accuracy, TNR, and TPR, clearly showed the high potential of the proposed method for performing classification tasks in bioinformatics and related disciplines. The initial value of δ as a ranking criterion was a key issue here for performing feature gene selection. In this paper, a flexible model for cancer gene expression classification and feature gene selection was proposed, which can adjust the parameters when using different datasets through cross validation to achieve the best result. The performance of the proposed method was compared to six classical methods, demonstrating that it could outperform existing methods in the identification of feature cancer genes. In conclusion, the proposed method is superior in accuracy and run-time for both two-class datasets and multiclass datasets, especially for those datasets with small samples. Furthermore, the results show that our method is computationally efficient. Also, the double-kernel learning may not be good at handling a super large scale of data. Future work could investigate computational aspects more in-depth on a large scale and use graph-based kernels to process gene networks.

## Additional files

**Additional file1:** Existing gene selection methods: a brief introduction. (DOCX 18 kb)

**Additional file 2:** SVM and KNN classifiers. (DOCX 18 kb)

**Additional file 3:** Dataset descriptions. (DOCX 16 kb)

**Additional file 4:** Further comparison for other datasets**. Table S1.** Average performance in KNN and SVM classifiers of DKBCGS and KBCGS (two classification). **Table S2.** Average performance in KNN and SVM classifiers of our method and KBCGS (multi-classification). **Table S3.** Performance of gene feature selection methods with KNN classifier (high) and SVM classifier (low) in two-class datasets. (DOCX 27 kb)

**Additional file 5:** top-50 gene expression**. Figure S1.** First fifty top-ranked gene expression level by different methods. The horizontal axis is the number of characteristic genes, the vertical axis is the gene expression level, and the black line represents the mean gene expression difference between the normal sample and the cancer sample. (PNG 596 kb)

### Abbreviations

ACC: Accuracy; DCBL: Diffuse large B-cell lymphoma; GRBF: Gaussian radial basis function; KNN: K-Nearest Neighbor; MRMR: Maximum relevance and minimum redundancy; PCA: Principal component analysis; POLY: Polynomial kernel function; RBF: Radial Basis Function; SRBCTs: Small round blue cell tumors; SVM: Support Vector Machine; TNR: True negative rate; TPR: True positive rate

### Acknowledgments

This work was partly supported by the Shandong Natural Science Foundation (ZR2015AM017) and the National Natural Science Foundation of China (Nos. 61877064). Matthias Dehmer thanks the Austrian Science Funds for supporting this work (project P 30031).

### Funding

This work was supported by The National Natural Science Foundation of China (Nos. 61877064), Shandong Natural Science Foundation (ZR2015AM017) and Austrian Science Funds (project P26142).

### Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

### Authors' contributions

SL and YZ proposed RBF-kernels. SL was a major contributor in programming. CX worked for the biological significance part of the manuscript. JL, BY, XL and MD jointly improved methods and applications. All authors contributed to analyzing gene expression data and writing manuscript, and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication
Not applicable.

### Competing interests
The authors declare that they have no competing interests.

## Publisher's Note


### Author details
[1]School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China. [2]Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [3]College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China. [4]Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr Campus, Steyr, Austria. [5]College of Computer and Control Engineering, Nankai University, Tianjin 300071, China. [6]Department of Mechatronics and Biomedical Computer Science, UMIT, Hall in Tyrol, Austria.

Received: 19 April 2018 Accepted: 26 September 2018
Published online: 29 October 2018

### References
1. Alizadeh A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403:503–11.
2. 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. 1999;96:6745–50.
3. Zhou X, Mao KZ. LS bound based gene selection for DNA microarray data. Bioinformatics. 2005;21(8):1559–64.
4. Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
5. Dhavala SS, Datta S, Mallick BK. Bayesian modeling of MPSS data: gene expression analysis of bovine Salmonella infection. Publ Am Stat Assoc. 2010;105(491):956–67.
6. Kira K, Rendell LA. A practical approach to feature selection. Int Workshop Mach Learn. 1992;48(1):249–56.
7. Chater N, Oaksford M. Information gain and decision-theoretic approaches to data selection: response to Klauer. Psychol Rev. 1999;106:223–7.
8. Peng H, Ding C, Long F. Minimum redundancy- maximum relevance feature selection. Bioinforma Comput Biol. 2005;3(2):185–205.
9. Saeys Y, et al. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
10. Blum A, Langley P. Selection of relevant features and examples in machine learning. Artif Intell. 1997;97:245–71.
11. Jacobs IJ, Skates SJ, Macdonald N. Screening for ovarian cancer: a pilot randomised controlled trial. Lancet. 1999;353(9160):1207–10.
12. Xiong M, et al. Biomarker identification by feature wrappers. Genome Res. 2001;11(11):1878–87.
13. Guyon I, et al. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
14. Kim D, Lee K, Lee D. Detecting clusters of different geometrical shapes in microarray gene expression data. Bioinformatics. 2005;7(1):3.
15. Duval B, Hao JK. Advances in metaheuristics for gene selection and classification of microarray data. Brief Bioinform. 2010;11(1):127–41.
16. Brenner S, Johnson M, Bridgham J. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000;18(6):630–4.
17. Bernhard S, Alexander JS. Learning with kernels. Cambridge: MIT Press; 2002.
18. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997;7(10):986.
19. Hanczar B, Dougherty ER. Classification with reject option in gene expression data. Bioinformatics. 2008;24(17):1889–95.
20. Chen H, Zhang Y, Gutman I. A kernel-based clustering method for gene selection with gene expression data. J Biomed Inform. 2016;62:12–20.
21. Smits GF, Jordan EM. Improved SVM regression using mixtures of kernels. Int Joint Conf Neural Netw. 2002;3:2785–90.
22. Scholkopf B, Mika S, Burges C, Knirsch P, Muller K, Ratsch G, Smola A. Input space versus feature space in kernel-based methods. IEEE Trans on Neural Networks. 1999;10:1000–17.
23. Phienthrakul T, Kijsirikul B. Evolutionary strategies for multi-scale radial basis function kernels in support vector machines. Conf Genet Evol Comput. 2005;14(7):905–11.
24. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 2002;8(1):68–74.
25. Rajkumar T, Sinha BN. Studies on activity of various extracts of Albiziaamara against drug induced gastric ulcers. Pharmacognosy J. 2011;3(25):73–7.
26. Yuan J, Yue H, Zhang M, Luo J. Transcriptional profiling analysis and functional prediction of long noncoding RNAs in cancer. Oncotarget. 2016; 7(7):8131–42.
27. Evgeniou TK. Learning with Kernel Machine Architectures. Ph.D. Dissertation. Cambridge: Massachusetts Institute of Technology; 2000. AAI0801924.
28. Frigui H, Nasraoui O. Simultaneous clustering and attribute discrimination. Proc Fuzzieee. 2000;1:158–63.
29. Merico D, Isserlin R, Bader GD. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. Methods Mol Biol. 2011;781:257–77.