Research Article

Comparison among dimensionality reduction techniques based on Random Projection for cancer classification

Haozhe Xie, Jie Li*, Qiaosheng Zhang, Yadong Wang

School of Computer Science and Technology, Harbin Institute of Technology, No. 92 Xidazhi Street, Harbin 150001, China

ARTICLE INFO

ABSTRACT

Random Projection (RP) technique has been widely applied in many scenarios because it can reduce high-dimensional features into low-dimensional space within short time and meet the need of real-time analysis of massive data. There is an urgent need of dimensionality reduction with fast increase of big genomics data. However, the performance of RP is usually lower. We attempt to improve classification accuracy of RP through combining other dimension reduction methods such as Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Feature Selection (FS). We compared classification accuracy and running time of different combination methods on three microarray datasets and a simulation dataset. Experimental results show a remarkable improvement of 14.77% in classification accuracy of FS followed by RP compared to RP on BC-TCGA dataset. LDA followed by RP also helps RP to yield a more discriminative subspace with an increase of 13.65% on classification accuracy on the same dataset. FS followed by RP outperforms other combination methods in classification accuracy on most of the datasets.

1. Introduction

Machine learning and data mining techniques have been widely applied in many areas (Berry and Linoff, 1997; Schmidt-Kastner et al., 2015; Han et al., 2011). The data has very high dimensionality in some of these application scenarios. For example, there are usually up to thousands of merchandises in a hypermarket. Obviously, the market basket data is very high-dimensional data. In text mining, documents are usually represented by a matrix whose dimensionality is equal to the vocabulary size. In these cases, the dimensionality of these matrices is very high due to either a considerable number of merchandises or a wealth of vocabulary. In addition to the high dimensionality, they are often sparse. For gene expression data with several thousands of genes and hundreds of samples, it can be considered as a 2D matrix with continuous value. Unlike the above-mentioned matrices, gene expression matrix is not sparse. High dimensionality may cause "curse of dimensionality" which results in inaccurate distance metrics and impact on classification precision. Traditional dimensionality reduction techniques include: Principle Component Analysis (PCA) (Jolliffe, 2002), Linear Discriminant Analysis (LDA) (Scholkopft and Mullert, 1999) and Feature Selection (FS) based on statistical test (Guyon and

Elisseeff, 2003; Li et al., 2007). PCA finds dimensions with maximum covariance in an unsupervised way, while LDA tries to seek the best linear discriminant function in a supervised way. FS is based on statistical test that selects features that are instructive in classification. Wrapper method for feature selection Kohavi and John (1997) trains a new model for classification for each subset, which can usually provide the best performing set for that particular type of model. In spite of this fact, direct use of PCA, LDA and FS may be problematic because these methods require impracticably large computational resources and cannot meet the needs of real-time processing. To address this problem, Random Projection (RP) was purposed in Dasgupta (2000), which projects original data onto a low-dimensional subspace randomly that the discriminative information can be approximately retained, and have been developing rapidly these years (Zhao and Mao, 2015; Zhao et al., 2016; Geppert et al., 2015). RP has attracted great deal of attention and has been successfully applied to solve large-scale or high-dimensional data analytics, such as classification (Zhao and Mao, 2015; Arriaga et al., 2015; Cannings and Samworth, 2015), clustering (Zhao et al., 2016; Tasoulis et al., 2014; Fern and Brodley, 2003), regression (Geppert et al., 2015; Maillard and Munos, 2012), manifold learning (Alavi et al., 2014; Freund et al., 2007), and information retrieval (Thaper et al., 2002). Compared to PCA, LDA and FS, RP is much less expensive in computational cost. However, RP may not capture task-related information because latent space is generated without considering the structure of original data, and accuracy of

* Corresponding author.
E-mail address: jieli@hit.edu.cn (J. Li).

classifiers based on RP is usually low. So we try to combine RP with other methods to reduce data dimensionality with the purpose of achieving a balance between computational complexity and classification accuracy.

In this paper, we combined RP with PCA, LDA and FS to find a low-dimensional subspace with better discriminative features to classify breast cancer datasets which have a huge amount of genes and small size samples. This paper is organized as follows. In Section 2, the related work is briefly introduced. The details of methods and experiments are described in Section 3 and 4 respectively. Finally, conclusion and future work is given in Section 5.

2. Related work

Assuming that the data is represented by a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n and d are the number of samples and dimension of dataset, and the dimension after reduction is k, where $k \ll d$. Dimensionality reduction maps original set of n d-dimensional observations to a k-dimensional subspace:

$$\hat{\mathbf{X}}_{n \times k} = \mathbf{X}_{n \times d} \mathbf{W}_{d \times k} \tag{1}$$

where $\hat{\mathbf{X}}_{n \times k}$ indicates the subspaces after reduction, and $\mathbf{W}_{d \times k}$ is the linear transformation matrix. Linear dimensionality reduction approaches are performed by their ways with different transformation matrix $\mathbf{W}$. In the following subsections, four dimensionality reduction methods used in this paper are briefly introduced.

2.1. Principle component analysis

PCA finds dimensions with maximum covariance in an unsupervised way (Jolliffe, 2002). It is mathematically defined as an orthogonal linear transformation to convert possibly correlated variables into values of linearly uncorrelated variables. The optimization problem for PCA can be summarized as follows:

$$\mathbf{W}^* = \mathrm{argmax}_{\mathbf{W} \in \mathbb{R}^{d \times k}} \left| \frac{\mathbf{W}^T \mathbf{S} \mathbf{W}}{\mathbf{W}^T \mathbf{W}} \right| \tag{2}$$

where $\mathbf{S}$ can be defined as follows:

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{3}$$

where $\mathbf{x}_i$ is a vector of d dimensions and $\bar{\mathbf{x}}$ represents the mean vector of all samples.

2.2. Linear discriminant analysis

LDA tries to seek the best linear discriminant function in a supervised way (Scholkopft and Mullert, 1999). Compared to PCA, LDA explicitly attempts to model the difference between the classes of data. LDA projects original data onto subspace which has sufficient discriminant power. The dimension of the subspace is restricted to be less than the number of classes. In particular, the dimension of subspace for binary classification problem is 1.

The optimization problem for LDA is described as follows:

$$\mathbf{W}^* = \mathrm{argmax}_{\mathbf{W} \in \mathbb{R}^{d \times k}} \left| \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}} \right| \tag{4}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are given as follows:

$$\begin{cases} \mathbf{S}_B = \sum_{c=1}^{Nc} n_c (\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^T \\ \mathbf{S}_W = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_{c_i})(\mathbf{x}_i - \bar{\mathbf{x}}_{c_i})^T \end{cases} \tag{5}$$

where N_c and n_c are the number of classes and the number of samples in cth class respectively, $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_c$ denote the mean vector of all samples and the cth class correspondingly, and c_i represents the label of the ith sample.

2.3. Feature selection

There are several statistical tests to select features (Li et al., 2007), such as rank sum test (Wilcoxon, 1945), t-test (Student, 1908) and so on. In the paper these methods are called feature selection methods. Here we take t-test as a representative to study the performance of FS methods derived from statistical test. The t-test assesses whether the means of two groups are statistically different from each other. Normalized distance between two classes of samples can be obtained using the sample mean values: m_c, m_t and the sample variances s_c^2 and s_t^2:

$$t = \frac{m_c - m_t}{\sqrt{\frac{s_c^2}{n_c} - \frac{s_t^2}{n_t}}} \tag{6}$$

where n_c and n_t denote the number of two classes of samples respectively, $m = \frac{1}{n} \sum_i x_i$ and $s^2 = \frac{1}{n} \sum_i (x_i - m)^2$ are the mean and variance of each sample respectively. Therefore, t follows a t-test with f degree of freedom:

$$f = \frac{[(s_c^2/n_c) + (s_t^2/n_t)]^2}{\frac{(s_c^2/n_c)^2}{n_c - 1} + \frac{(s_t^2/n_t)^2}{n_t - 1}} \tag{7}$$

Two classes of samples are considered significantly different if t exceeds a certain threshold of the selected confidence interval.

The ith sample is denoted by $x_i (i = 1, 2, \ldots, d)$, and the corresponding label is y_i, where $y_i \in \{0, 1\}$. To simplify the problem, we assume $x_i (i = 1, 2, \ldots, d)$ is independent. In this paper, k features with significant differences are selected to build a classifier. The optimal transformation vector $\mathbf{W}$ in Eq. (1) can be formed as follows:

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \ldots & w_{1,k} \\ w_{2,1} & w_{2,2} & \ldots & w_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \ldots & w_{d,k} \end{bmatrix} \tag{8}$$

where

$$w_{i,j} = \begin{cases} 1 & i\text{-th feature is selected} \\ 0 & i\text{-th feature is not selected} \end{cases} \tag{9}$$

2.4. Random Projection

In Bingham and Mannila (2001) and Li et al. (2006), RP has been proposed to address the burdensome computation time in dimensionality reduction. In RP, transformation matrix $\mathbf{W}$ is randomly generated. Experiment results (Bingham and Mannila, 2001; Li et al., 2006) have revealed that RP is computationally less expensive with a little distortion of original data. On the other hand, RP may not capture intrinsic structure underlying original data because $\mathbf{W}$ is generated without considering the structure of data.

As suggested in Li et al. (2006), the linear transform matrix $\mathbf{W}$ is sampled as follows:

$$w_{i,j} = \sqrt{c} \begin{cases} 1 & \text{with prob.} \dfrac{1}{2c} \\ 0 & \text{with prob.} 1 - \dfrac{1}{c} \\ -1 & \text{with prob.} \dfrac{1}{2c} \end{cases} \tag{10}$$

where c is set to $\sqrt{d}$.

3. Combination methods based on Random Projection

In this section, RP is combined with other methods, such as PCA, LDA and FS to improve the performance of RP.

3.1. Random Projection + Principle Component Analysis (RP + PCA)

PCA seeks direction with maximum variance. We wonder whether RP based on PCA can improve classification accuracy.

Assume that matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is consisted of n samples and d dimension. Firstly, RP is used to map original data to a subspace $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ by a random matrix $\mathbf{W}_1 \in \mathbb{R}^{d \times k_1}$ according to Eq. (10). Then PCA projects $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ to $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ by a matrix $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ from Eq. (2).

3.2. Random Projection + Linear Discriminant Analysis (RP + LDA)

Here we want to know whether LDA is benefit for RP. RP is combined with LDA which is a supervised approach to find the best linear discriminant function.

Suppose data is described by a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. First RP maps original data to a subspace $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ by a random matrix $\mathbf{W}_1 \in \mathbb{R}^{d \times k_1}$ announced in Eq. (10). Then LDA coverts $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ to $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ by a linear transformation matrix $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ according to Eq. (4).

3.3. Random Projection + Feature Selection (RP + FS)

In subspaces generated by RP, some dimensions do not have significant differences, and this may affect the accuracy in classification. We believe FS can help to filter them.

According to Eq. (10), a subset $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ with k_1 features is generated in a random way. Later, FS projects data onto $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ by a matrix $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ declared in Eq. (8).

3.4. Feature Selection + Random Projection (FS + RP)

In this method, FS is used before RP to filter dimensions that are similar between two classes of samples in original dataset. The preprocessing conducted by FS may help RP yield more discriminant subspaces.

Matrix $\mathbf{W}_1 \in \mathbb{R}^{d \times k_1}$ with k_1 features is firstly calculated by FS with a matrix described in Eq. (8). Another matrix $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ is calculated using Eq. (10) in order to obtain a subspace $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ by RP.

3.5. Feature Selection + Random Projection + Linear Discriminant Analysis (FS + RP + LDA)

Here three methods: FS, RP and LDA are combined to reduce dimensionality. Compared with FS + RP, LDA is added to study whether LDA is useful to improve the subspace generated by FS + RP.

This method includes three steps: first of all, FS transforms original data into $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ with a matrix $\mathbf{W}_1 \in \mathbb{R}^{d \times k_1}$ stated in Eq. (8). Secondly, RP is employed to generate a linear transformation matrix $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ whose values are random assigned and maps data to a subspace $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ with k_2 dimension. Finally, LDA is taken to form $\hat{\mathbf{X}}_3 \in \mathbb{R}^{n \times k_3}$ with a matrix $\mathbf{W}_3 \in \mathbb{R}^{k_2 \times k_3}$ determined in Eq. (4).

3.6. Feature Selection + Random Projection + Post Feature Selection (FS + RP + PFS)

In the method, we plan to employ a new method called "Post Feature Selection" (PFS) to select features after FS + RP in order to find a set of features which have sufficient discriminative power.

PFS picks column vectors out of $\mathbf{X}$ that have the maximum classification accuracy. y_i denotes the label of the ith sample x_i of $\mathbf{X}$. Training and testing datasets include m and n samples respectively are randomly selected from $\mathbf{X}$. A Support Vector Machine (SVM) with radial basis kernel function is used to evaluate the classification accuracy for each dimension, where top k dimensions with higher accuracy are selected.

This method firstly derives a subspace $\hat{\mathbf{X}}_1 \in \mathbb{R}^{n \times k_1}$ using a matrix $\mathbf{W}_1 \in \mathbb{R}^{d \times k_1}$ announced in Eq. (8). A subset $\hat{\mathbf{X}}_2 \in \mathbb{R}^{n \times k_2}$ with k_2 features is calculated in a random way with $\mathbf{W}_2 \in \mathbb{R}^{k_1 \times k_2}$ in the next stage. Finally, PFS is adopted to generate $\mathbf{W}_3 \in \mathbb{R}^{k_2 \times k_3}$ and corresponding subspace $\hat{\mathbf{X}}_3 \in \mathbb{R}^{n \times k_3}$, where k_3 is the number of features selected in PFS.

4. Experiments

4.1. Datasets and experimental environment

In this section, the performance of methods described above was tested and compared on three breast cancer gene expression data (Network et al., 2012; Wang et al., 2005; Hatzis et al., 2011), which can be downloaded from http://tcga-data.nci.nih.gov/tcga and http://www.ncbi.nlm.nih.gov/geo. Breast cancer from TCGA (BC-TCGA) (Network et al., 2012) consists of 17,814 genes and 590 samples (including 61 normal tissue samples and 529 breast cancer tissue samples). GSE2034 (Wang et al., 2005) includes 12,634 genes and 286 breast cancer samples (including 107 recurrence tumor samples and 179 no recurrence samples). GSE25066 (Hatzis et al., 2011) has 492 breast cancer samples available (including 100 pathologic complete response (PCR) samples and 392 residual disease (RD) samples) and 12,634 genes. Some missing values in these datasets were filled with gene mean values of each group and expression values of each gene were processed by Z-score. The above methods were implemented with scikit-learn,[1] which is one of the most popular packages for machine learning in Python. The environments for all methods performed in our experiments are as follows: a Linux PC running with Intel Xeon E5504 (2.00 GHz) processor and 24 GB RAM. In order to test the classification accuracy and running time of different methods, a SVM was employed to classify two classes of samples. Thus, methods' running time mainly includes two parts: one is the time spending on reducing dimensionality; the other is the time used for training a SVM.

---

[1] http://scikit-learn.org/stable/.

Table 1
Highest classification accuracy (%) on three gene expression datasets.

Methods	BC-TCGA		GSE2034		GSE25066	
	Training set	Testing set	Training set	Testing set	Training set	Testing set
RP	100.00	86.23	97.85	59.59	100.00	66.90
FS + RP	99.95	**98.97**	99.88	61.25	96.48	**71.07**
FS + RP + LDA	99.93	98.67	80.71	60.56	87.81	69.56
FS + RP + PFS	99.97	98.95	99.99	**61.55**	68.50	67.06
RP + FS	100.00	92.70	100.00	60.89	99.98	68.65
RP + PCA	100.00	76.70	100.00	54.25	100.00	60.44
RP + LDA	98.68	98.00	73.65	58.87	80.23	69.21

Bold face indicates the highest classification accuracy on test data.

4.2. Experimental procedure

Here we take RP + PCA as an example to introduce the detailed procedure of the whole experiment on BC-TCGA. First, we get training and testing sets. Training sample set includes 30 normal and 30 tumor tissue samples randomly selected from BC-TCGA without replacement, and testing sample set includes 30 normal and 30 tumor tissue samples randomly selected from remaining samples. Then, the training dataset is mapped into a subspace $\hat{\mathbf{X}}'_{\text{Train}}$ by a random matrix $\mathbf{W}_{\text{RP}}$. Then, a linear matrix $\mathbf{W}_{\text{PCA}}$ derived from PCA projects $\hat{\mathbf{X}}'_{\text{Train}}$ onto another subspace $\hat{\mathbf{X}}_{\text{Train}}$. Next, $\mathbf{W}_{\text{RP}}$ and $\mathbf{W}_{\text{PCA}}$ are applied to test dataset to obtain $\hat{\mathbf{X}}_{\text{Test}}$. Finally, a SVM with radial basis kernel function is trained and tested on $\hat{\mathbf{X}}_{\text{Train}}$ and $\hat{\mathbf{X}}_{\text{Test}}$ respectively. When the number of spatial dimension changes, classification accuracy and running time of SVMs also change for all methods described above. In order to investigate the effect of change in dimension, we conducted a sequence of experiments in different dimensions: 10, 50, 100, 200, 400, 800, 1000, 1250, 1500, 1750, 2000, 2500, 3000, 4000, and 5000. The above procedure is repeated 100 times. Other methods also follow a similar process on these datasets, except for the difference that training and testing sets are consisted of 50 positive samples and 50 negative samples. Positive samples stand for no recurrence samples in GSE2034 and PCR samples in GSE25066, while negative samples refer to recurrence samples in GSE2034 and RD samples in GSE25066. Finally, average values of classification accuracy and corresponding running time are obtained.

4.3. Experimental results

The experiment results on these gene expression datasets are shown in Table 1 and Fig. 4. Classification accuracy of different methods on the three datasets is shown in Table 1. It can be seen that accuracy of dimensionality reduction methods on these datasets is different. On the first dataset all methods get high classification accuracy on testing set, while on the other two datasets obtain relatively low accuracy. This indicates that two classes of samples in GSE2034 and GSE25066 seem to be more difficult to classify.
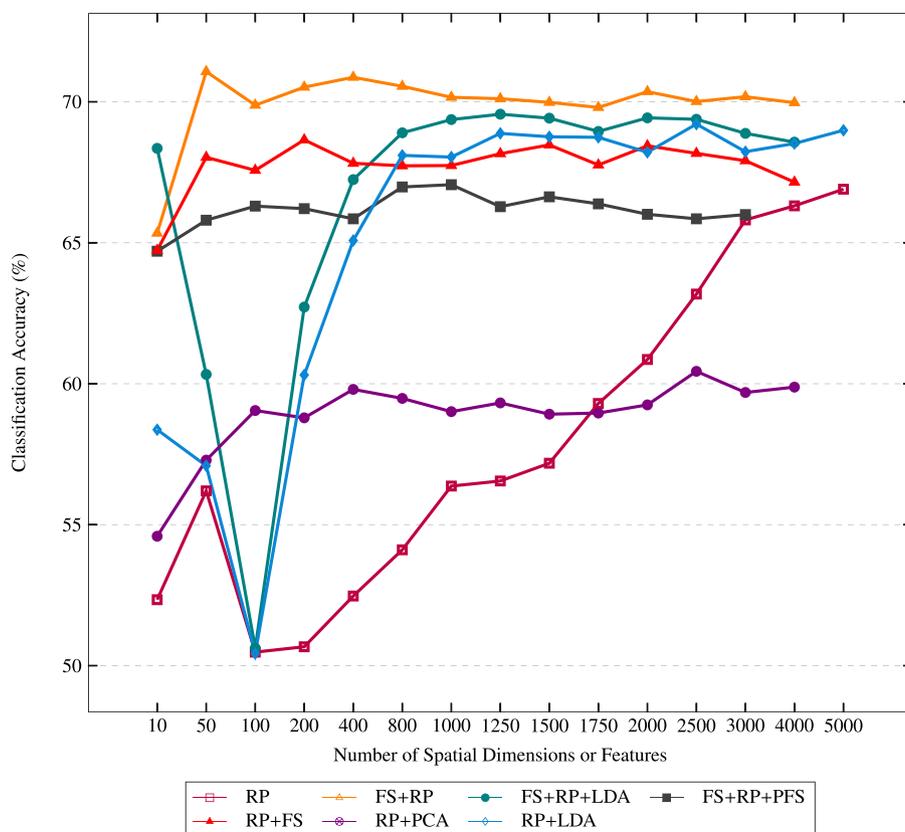

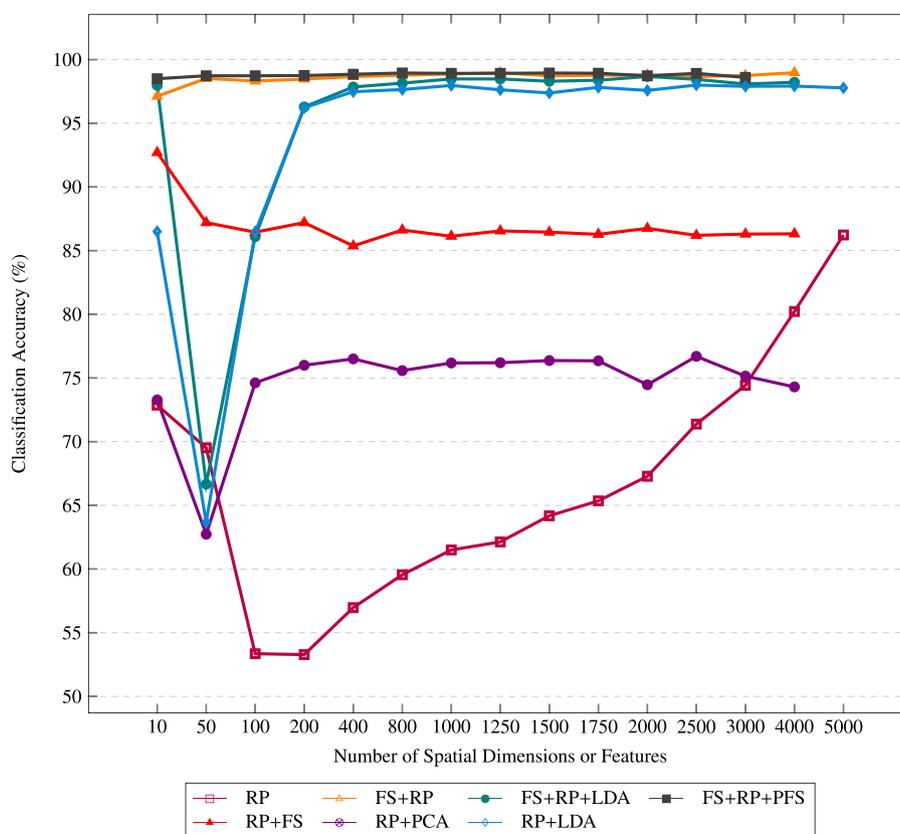**Fig. 1.** Classification accuracy of different methods on the testing set of GSE25066.

Fig. 2. Classification accuracy of different methods on the testing set of BC-TCGA.

The classification accuracy of RP and FS + RP is 86.23% and 98.97% on the first test data respectively, where the latter are significantly higher. On the other two datasets, the SVM based on FS + RP also has higher classification accuracy than the SVM based on RP, where the classification accuracy increases by 2.79% and 6.23% respectively. The accuracy of SVM based on FS + RP + LDA and FS + RP + PFS is a bit lower than that of FS + RP on the first test data, with an accuracy of 98.67% and 98.95% respectively, while on GSE2034, FS + RP + PFS performs a little better than FS + RP with an accuracy of 61.55%. SVM based on FS + RP obtain highest classification accuracy of 71.05% on the third test dataset.

LDA can also help RP to find a latent space, where the classification accuracy of SVM based on RP + LDA on BC-TCGA and GSE25066 is 98.00% and 69.21% respectively, while the accuracy on GSE2034 is 58.87% which is even worse than that of RP. Recent studies have indicated that classification accuracy of LDA is afflicted by two classes of samples with small inter class distance (Huang et al., 2002). For this reason, RP + LDA deteriorates subspace generated by RP. And experiments in the next subsection also prove this point. In contrast, methods combined with FS still lead to good classification accuracy even distance between classes is very small. Due to this, RP + FS achieves better classification accuracy of 60.89% on GSE2034 compared to RP + LDA. On all datasets, classification accuracy of SVM based on RP is lower than that of RP + FS, because RP projects original data to subspace where samples cannot be effectively divided into two groups in some dimensions (named "invalid dimensions") and FS filters out features from invalid dimensions after RP.

The accuracy of PCA and RP + PCA on the testing set of BC-TCGA is 77.31% and 76.70% respectively, which is significantly lower than 86.23% of RP. However, the accuracy of SVM based on RP + PCA is much lower than that of FS + RP, RP + FS and RP + LDA. Because RP

and PCA are both unsupervised approaches and features generated by them may not yield better classification performance.

Classification accuracy changes with the number of selected features (or spatial dimensions). Classification accuracy of the different methods on the testing set of GSE25066 is shown in Fig. 1. The classification accuracy of FS + RP, FS + RP + PFS, RP + PCA and RP + FS has small changes with the change of dimension, as relatively large. There is a steady increase in accuracy of RP when the dimension goes up. According to Johnson–Lindenstrauss (JL) lemma, pairwise distances are preserved more preciously with the dimension of subspace rises. However, for dimension below the minimal number of components to guarantee with good probability (i.e. 148 for GSE25066), JL lemma is not suitable for this case and the classification accuracy of these subspaces is determined by intrinsic structure of data. Moreover, there is a sharply increase in the accuracy of FS + RP + LDA and RP + LDA when the dimension goes up. Compared with other methods, FS + RP has highest accuracy (around 70%) in almost all dimensions, which achieves a significant improvement of 40% in dimension equals to 100 compared to RP. It is interesting that the classification accuracy of FS + RP, RP + FS, and FS + RP + PFS remains 65–70% even in lower dimensions. Fig. 1 also reveals that the accuracy of RP + PCA is relatively low (about 60%) and has small changes with the increase of dimensions. This indicates that PCA cannot improve the performance of RP. The similar results also obtained on other two datasets (Figs. 2 and 3).

It is hard to compare running time since the dimension of classifiers based on different methods is different when the highest accuracy is obtained. As shown in Fig. 1, FS + RP, FS + RP + PFS, RP + FS, and RP + PCA have relatively stable classification accuracy when dimension is 150, while FS + RP + LDA, RP + LDA and RP have relatively low classification accuracy. And dimension equals to 1000, most methods except RP obtain stable classification accuracy.
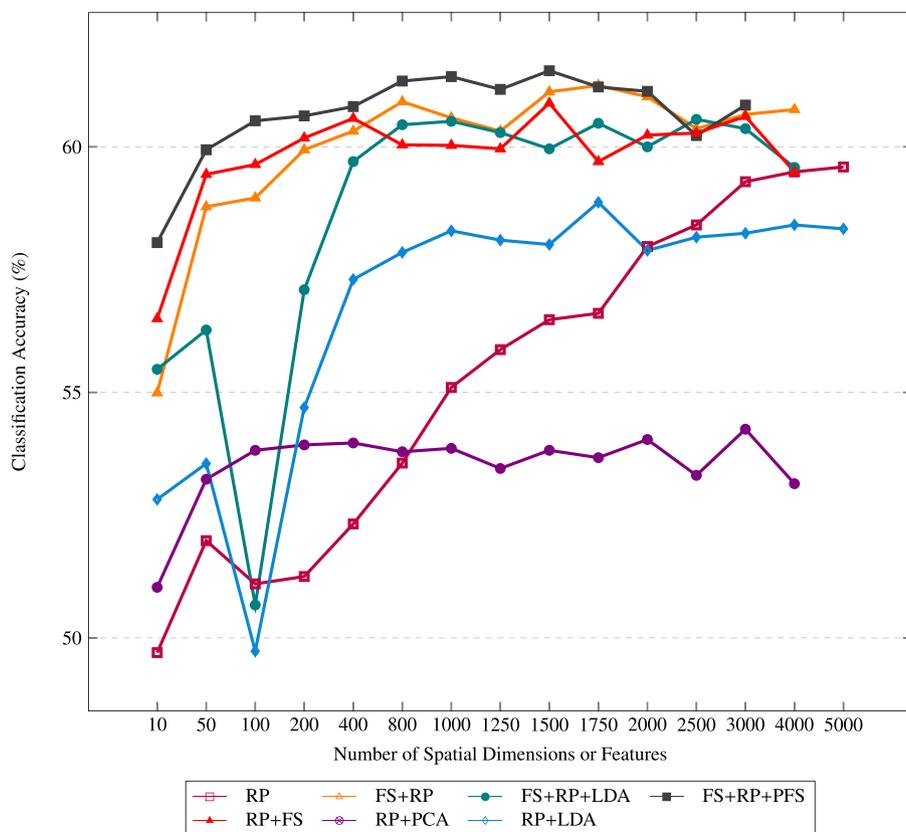
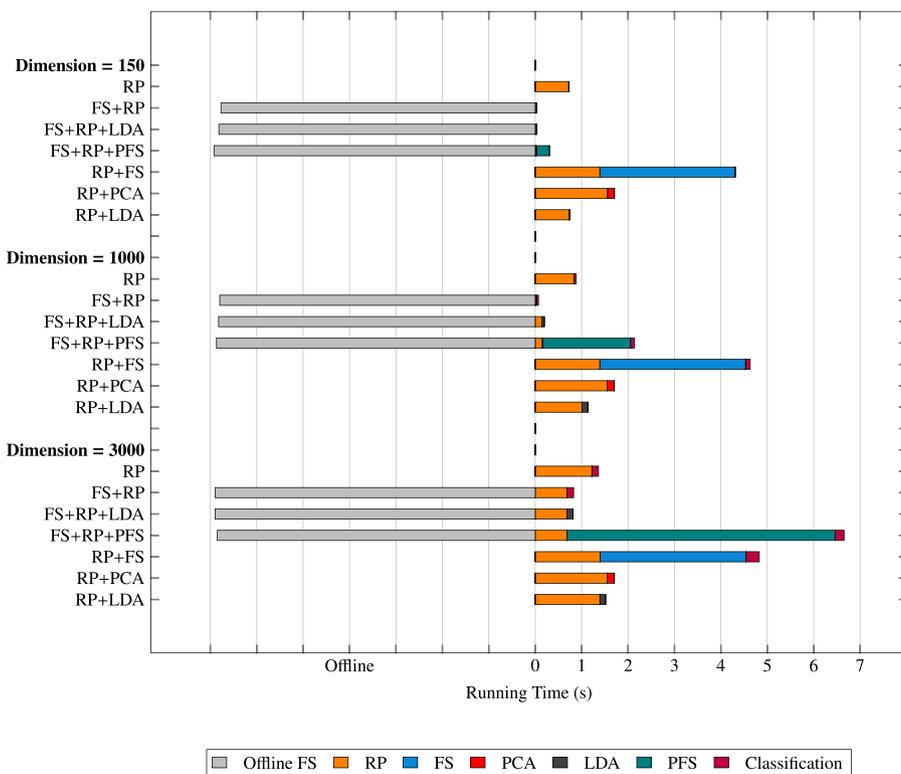**Fig. 3.** Classification accuracy of different methods on the testing set of GSE2034.

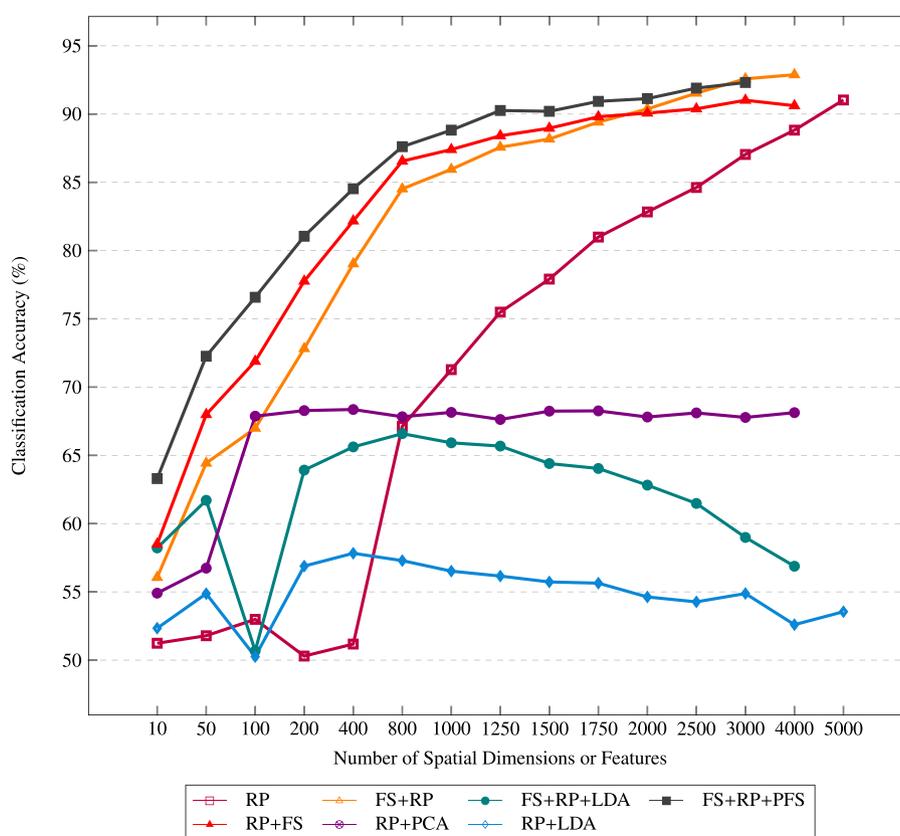**Fig. 4.** Running time of different methods on GSE25066.

Fig. 5. Classification accuracy of different methods on the testing set of SData.

When dimension rise to 3000, all methods have stable classification accuracy. Here we compared the running time in three different dimensions: 150, 1000, and 3000 on one of datasets (GSE25066), which is displayed in Fig. 4, where consuming time of each stage in dimensional reduction methods was marked with different colors.

In many scenarios, prior knowledge and empirical data can be used in feature selection for offline processing. For example, microarray data can be preprocessed. Fig. 4 shows that FS + RP has the least online processing time among all methods. It indicates that offline preprocessing can not only improve classification accuracy but also greatly accelerate the speed of online processing. Furthermore, RP + LDA obtains a more discriminant subspace for linear spreadable data than RP consuming almost the same time. It is clear that the running time rises with the dimensions, and the running time of FS + RP + PFS increases the most among all methods.

4.4. Verification

To further validate the performance of different methods, we repeat the above experiments on a simulation dataset: SData, which includes 100 positive samples and 100 negative samples with 10,000 features, and each feature in SData follows normal distributions: $N(0, 0.1)$ and $N(0 \pm r, 0.1)$ for positive and negative samples respectively, here $r \in [-0.125, 0.125]$. The performance of all methods on SData is listed in Table 2. Fig. 5 displays the accuracy of different methods with the change of dimension on SData. There is a rapid increase in RP with the increase of dimension from 100 to 800. Compared with RP, FS + RP, FS + RP + PFS and RP + FS yield better classification accuracy in low-dimensional subspace. As mentioned above, LDA leads to terrible performance on two classes of samples having similar mean values. Therefore, RP + LDA and FS + RP + LDA have low accuracy on SData. Overall these methods show the similar performance on the real and the simulated datasets.

Table 2
Highest classification accuracy (%) on SData.

Methods	SData	
	Training set	Testing set
RP	100.00	91.03
FS + RP	100.00	**92.88**
FS + RP + LDA	88.74	66.59
FS + RP + PFS	100.00	92.31
RP + FS	100.00	91.01
RP + PCA	100.00	68.36
RP + LDA	74.87	57.83

5. Conclusions

In this paper, we compared performance of different techniques based on RP, and experimental results proved that the classification accuracy of RP can be improved by combining with other dimensionality reduction methods, such FS or LDA. However, it did not yield better classification accuracy combining RP with PCA. FS followed by RP outperforms other methods in classification accuracy on most of the datasets with relatively lower computation cost.

Conflict of interest

The authors declare that they have no competing interests.

Funding information

References

Alavi, A., Wiliem, A., Zhao, K., Lovell, B.C., Sanderson, C., 2014. Random projections on manifolds of symmetric positive definite matrices for image classification. In: 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 301–308.

Arriaga, R.I., Rutter, D., Cakmak, M., Vempala, S.S., 2015. Visual categorization with random projection. Neural Comput.

Berry, M.J., Linoff, G., 1997. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Bingham, E., Mannila, H.,2001. Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 245–250.

Cannings, T.I., Samworth, R.J., 2015. Random projection ensemble classification. arXiv:1504.04595.

Dasgupta, S.,2000. Experiments with random projection. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 143–151.

Fern, X.Z., Brodley, C.E., 2003. Random projection for high dimensional data clustering: a cluster ensemble approach. In: ICML, vol. 3, pp. 186–193.

Freund, Y., Dasgupta, S., Kabra, M., Verma, N., 2007. Learning the structure of manifolds using random projections. In: Advances in Neural Information Processing Systems., pp. 473–480.

Geppert, L.N., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C., 2015. Random projections for Bayesian regression. Stat. Comput., 1–23.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques. Elsevier, New York, USA.

Hatzis, C., Pusztai, L., Valero, V., Booser, D.J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H., et al., 2011. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305 (18), 1873–1881.

Huang, R., Liu, Q., Lu, H., Ma, S.,2002. Solving the small sample size problem of LDA. In: 16th International Conference on Pattern Recognition, 2002. Proceedings, vol. 3. IEEE, pp. 29–32.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library, Hoboken, NJ, USA.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1), 273–324.

Li, J., Tang, X., Zhao, W., Huang, J., 2007. A new framework for identifying differentially expressed genes. Pattern Recognit. 40 (11), 3249–3262.

Li, P., Hastie, T.J., Church, K.W.,2006. Very sparse random projections. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 287–296.

Maillard, O.-A., Munos, R., 2012. Linear regression with random projections. J. Mach. Learn. Res. 13 (1), 2735–2772.

Network, C.G.A., et al., 2012. Comprehensive molecular portraits of human breast tumours. Nature 490 (7418), 61–70.

Schmidt-Kastner, R., Wilneff, J., Nguyen, C., McKain, K., Cruz, J.D., Steinbusch, H., Rutten, B., 2015. Data-mining of genomic data in Parkinson's disease for links to hypoxia-response and vascular factors (P6. 060). Neurology 84 (14 Suppl.), 6–060.

Scholkopft, B., Mullert, K.-R., 1999. Fisher discriminant analysis with kernels. Neural Netw. Signal Process. IX 1 (1), 1.

Student, 1908. The probable error of a mean. Biometrika, 1–25.

Tasoulis, S., Cheng, L., Valimaki, N., Croucher, N.J., Harris, S.R., Hanage, W.P., Roos, T., Corander, J.,2014. Random projection based clustering for population genomics. In: 2014 IEEE International Conference on Big Data (Big Data). IEEE, pp. 675–682.

Thaper, N., Guha, S., Indyk, P., Koudas, N.,2002. Dynamic multidimensional histograms. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. ACM, pp. 428–439.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al., 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365 (9460), 671–679.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80–83.

Zhao, K., Alavi, A., Wiliem, A., Lovell, B.C., 2016. Efficient clustering on Riemannian manifolds: a kernelised random projection approach. Pattern Recognit. 51, 333–345.

Zhao, R., Mao, K., 2015. Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space. IEEE Comput. Intell. Mag. 10 (3), 30–41.