Neural Comput & Applic (2017) 28:3047–3059
DOI 10.1007/s00521-016-2737-2

Unsupervised feature selection based on decision graph

Jinrong He[1,2] · Yingzhou Bi[2] · Lixin Ding[3] · Zhaokui Li[4] · Shenwen Wang[5]

Received: 6 April 2016 / Accepted: 23 November 2016 / Published online: 2 December 2016

Abstract In applications of algorithms, feature selection has got much attention of researchers, due to its ability to overcome the curse of dimensionality, reduce computational costs, increase the performance of the subsequent classification algorithm and output the results with better interpretability. To remove the redundant and noisy features from original feature set, we define local density and discriminant distance for each feature vector, wherein local density is used for measuring the representative ability of each feature vector, and discriminant distance is used for measuring the redundancy and similarity between features. Based on the above two quantities, the decision graph score is proposed as the evaluation criterion of unsupervised feature selection. The method is intuitive and simple, and its performances are evaluated in the data classification experiments. From statistical tests on the averaged classification accuracies over 16 real-life dataset, it is observed that the proposed method obtains better or comparable ability of discriminant feature selection in 98% of the cases, compared with the state-of-the-art methods.

✉ Jinrong He
hejinrong@nwafu.edu.cn

[1] College of Information Engineering, Northwest A & F University, Yangling, Shaanxi 712100, China

[2] Science Computing and Intelligent Information Processing of Guangxi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning 530023, China

[3] School of Computer, Wuhan University, Wuhan, Hubei 430072, China

[4] School of Computer, Shenyang Aerospace University, Shenyang, Liaoning 110136, China

[5] School of Information Engineering, Hebei Dizhi University, Shijiazhuang, Hebei 050031, China

Keywords Feature selection · Decision graph · Local density · Discriminant distance

1 Introduction

Feature learning is an important step in machine learning and data mining, which has been widely applied in many big data analysis domains, such as gene microarray data, text data and image sequences in video processing. There are two ways to generate features from data samples: feature transformation and feature selection. New features generated from feature transformation are some combinations of original features, while new features from feature selection are just a subset of original features. As a dimensionality reduction method, feature selection can effectively remove redundant features which are irrelevant to data classification task and retain a small number of key features, which not only reduce computational complexity of data classification or clustering, but also improve accuracy of machine learning algorithms. Compared with feature transformation, the selected features have better explanations, since they are the subset of original high-dimensional features, which have specific physical meanings. Therefore, it has obtained many attentions from researchers.

From the point of view of combinational optimization, feature selection is a NP-hard problem [1]. Essentially, feature selection aims to rank features with their importance and then select the most important features for subsequent data analysis. Therefore, feature selection methods are derived from different feature importance evaluation criteria. According to independence between the feature generation process and follow-up training process of learning model, feature selection methods can be divided

into two categories: packaging and filtering. According to whether labels of samples are used in feature selection process, it can be divided into supervised and unsupervised methods. Generally, packaging methods are commonly used in supervised feature selection algorithm design, and filtering methods are commonly used in unsupervised feature selection algorithm design.

Supervised methods usually select features based on correlations between features and labels. For example, according to correlations, Relief proposed by Kira et al. [2] firstly puts different weight for each feature and then deletes irrelevant features with a preset threshold. The original Relief algorithm only applies to binary classification problem; in order to deal with multi-classification problems, Kononeill et al. proposed an improved method named Relief-F [3]. FOCUS-SF [4] finds the minimum subset of features that is consistent with labels by searching all feature subsets exhaustively. Due to the high computational complexity, FOCUS-SF is not suitable for high-dimensional feature selection. Correlation-based feature selection (CFS) [5] takes the correlation between feature and feature or feature and class as evaluation index of feature importance, and then, the optimal feature subset can be found under a specific search strategy. Fast correlation-based filter (FCBF) [6] introduces concept of dominant correlation to evaluate feature importance. Since feature selection can be seen as a special kind of subspace learning, Nie et al. [7] formulate the feature selection process as trace ratio criterion-based graph embedding optimization problem, in which each column of projection matrix from high-dimensional data to low-dimensional representations is constrained as one-hot code.

In many applications, there are a large number of unlabeled data, since it is time-consuming to obtain sample labels, so unsupervised feature selection methods have wider range of applications. For example, Wei et al. [8] propose maximum overall dependence-based forward orthogonal search algorithm (FOS-MOD) for feature selection, which takes the representative ability of one feature to others as feature importance evaluation index. The similarity between two features can be measure by squared correlation coefficient in FOS-MOD, and then, the optimal features can be obtained by selecting features with largest squared correlation coefficient step by step. Li et al. [9] take sample margin and hypothesis margin of each feature as feature importance evaluation index, respectively, and then select features based on sequential backward method, which can be classified by support vector machine (SVM). Recently, hybrid feature selection algorithms have gained great importance in terms of timeliness. For example, Brahim et al. [29] proposed a feature selection method to design an intelligent assistance sleep scoring system. Based on instance learning, Sen et al. [30] proposed a filter wrapper method for feature selection by cooperative subset search.

Information theory is a powerful mathematical tool for description of the interaction between variables. Based on some concepts of information theory, such as mutual information and entropy, it has generated a lot of unsupervised feature selection algorithms. For example, Dash et al. [10] proposed a consistency measure of feature subset for any search strategy, which assumes that the categories of samples with same feature subset should be the same. In addition, the concept of entropy is used to measure whether the dataset has obvious clusters, which can also be used as the feature importance measure [11]. Mitra et al. [12] proposed an unsupervised feature selection method based on maximum information compression coefficient, which is defined as the minimum eigenvalue of covariance matrix between two random variables. Since the projection direction corresponding to minimum eigenvalue is orthogonal to directions corresponding to the principal components, it can be used to measure dissimilarity between the two features vectors, and then, redundant features can be eliminated. Peng et al. [13] proposed a mutual information-based maximum statistical dependence criterion for incremental feature selection. For computational efficiency, maximum statistical dependence can be transformed into minimum redundancy maximum correlation (mRMR) model. The mRMR model has been successfully used for handwritten digital images feature selection [14]. Based on mutual information, Xu et al. [15] use minimum redundancy and maximum correlation to evaluate feature importance, where the correlation is the degree of dependence between a feature vector and its potential class, and redundancy is the degree of dependence between two features. Both of them can be measured by mutual information. Bandyopadhyay et al. [16] proposed an unsupervised feature selection method based on dense subgraph discovery, in which each feature vector can be viewed as a vertex of the graph, and the mutual information between feature vectors can be viewed as the weight of edge on the graph. After finding the dense subgraph, optimal features can be selected by clustering.

Since manifold can model the low-dimensional structure of datasets, it has been used in unsupervised feature selection. Because the intra-class samples are also locally nearby, Laplacian Score [17] uses the locality preserving ability of a feature to describe its importance. Based on spectral graph theory, Zhao and Liu [18] unify supervised and unsupervised feature selection into a framework. After the similarity between two samples is defined, the structure of dataset can be described as a graph, and then, the normalized graph cut can be used as the feature importance measure. In order to fully exploit the discriminant structure of datasets, Cai et al. [19] proposed a multi-cluster feature

selection method (MCFS) [19], in which the original high-dimensional dataset is firstly projected into low-dimensional space by spectral embedding; then, the linear dependence relationship between high-dimensional sample and its low-dimensional representation can be obtained by L_1 norm regularized regression problem, and the features corresponding to largest sparse representation coefficients are optimal features. Yang et al. [20] propose an unsupervised discriminant feature selection method (UDFS), in which samples are assumed to be linearly separable; then, according to linear relationship between a sample and its local label, the optimal discriminant feature subset can be obtained by maximizing the local interclass scatters and minimizing the local intra-class scatters, simultaneously minimizing $L_{2,1}$ norm of linear classification coefficient matrix. In addition, the label information of samples can be obtained through clustering, and then, the label information can guide the discriminant feature extraction, so Li et al. [21] proposed a nonnegative discriminant feature selection method (NDFS), which unified spectral clustering and feature selection into an optimization objective, the indicator matrix of cluster is constrained to be nonnegative. Similarly, Du et al. [22] proposed a local and global discrimination learning-based feature selection method (LGDFS), in which the weighted L_2 norm regularized regression models are optimized simultaneously locally and globally, and then, the optimal feature subset can be obtained from feature indicator matrix. Many manifold learning-based feature selection methods can be unified into a similarity preserving feature selection (SPFS) framework [23], which is equivalent to multivariate multi-output regression problem essentially. Different constraints, regularization conditions and optimization strategies result in different feature selection methods.

The integration of clustering and unsupervised feature selection can improve the discriminant ability of selected feature subset. Liu et al. [24] proposed a K-means-based feature selection method (KFS) for text clustering. After selecting different K and initialization samples to get different clustering results, the feature importance can be computed by χ^2 statistics on these clustering results, and then, the optimal feature subset can be obtained by ranking the sum of different feature importance computations. Similar to principal component analysis, principal feature analysis (PFA) [25] projects each feature vector into the subspace with maximum variance, in which all features are clustered by K-means, and then, the optimal feature subset can be obtained by the distance between the feature vector and its corresponding cluster center. Song et al. [26] proposed a clustering-based fast feature selection algorithm (FAST), which divides feature set into different clusters by minimum spanning tree-based clustering method, and then, the most representative features related to classification are selected from each cluster. Yan and Yang [27] proposed a sparse discriminant feature selection method (SDFS) by minimizing intra-class reconstruction residuals and maximizing interclass reconstruction residuals, in which the $L_{2,1}$ norm minimization can remove the redundant features effectively.

However, manifold learning methods rely heavily on data graph construction, and they are very sensitive to noises or corruptions. On the other hand, the features related to classification or clustering tasks are also correlated, so the most feature selection methods cannot reduce the redundancy of selected feature subset effectively. This paper proposed a decision graph-based feature selection (DGFS). Decision graph is a powerful tool for discovering clustering structure of feature set. The feature centered on each cluster is most representative and has minimum redundancy to others. Compared with other methods, DGFS has an intuitive principle and simple computation. The classification experiments on face datasets and UCI datasets show that DGFS can reduce the redundancy information contained in feature set effectively, and the selected features have a better ability of discriminant.

2 Decision graph-based feature selection

2.1 Problem formulation

For concise description, the observed samples are represented as a data matrix, i.e., n m-dimensional samples can be denoted as $X = (x_1, x_2, \ldots, x_n) \in R^{m\times n}$, where each column of X is an observed sample, and each row of X is a feature vector or attributes of observed samples. If the ith feature vector of dataset X is denoted as f_i, then the data matrix can also be rewritten as $X = (f_1; f_2; \ldots; f_m)$. Feature selection aims to select r features from m features, according to a specific feature importance evaluation criterion, such that the redundancy or correlation between these r features is small, and meanwhile, they can preserve most information contained in the original dataset.

2.2 Decision graph

In order to recognize most discriminant features in original high-dimensional data samples, the concept of decision graph [28] is introduced as follows.

Definition 1 Local density ρ_i on feature vector f_i is defined as

$$\rho_i = \sum_{j=1}^{m} \theta(d_c - d_{ij}) \tag{1}$$

where d_{ij} is the Euclidean distance between feature vectors f_i and f_j, and $d_c > 0$ is a preset threshold or truncated distance, and $\theta(\cdot)$ is an indicator function which can be defined as:

$$\theta(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases}$$

According to Definition 1, local density on each feature vector is the number of features contained in a neighborhood ball which is centered in the feature vector with a positive radius. Large ρ_i indicates dense feature distribution in the neighborhood of feature vector f_i. Therefore, it is reasonable to use ρ_i as descriptor for distribution of feature set. In applications, any other similar definitions of local density can be used, such as

$$\rho_i = \sum_{j=1}^{n} e^{\left(\frac{d_{ij}}{d_c}\right)^2} \tag{2}$$

Definition 2 Discriminant distance δ_i on feature vector f_i is defined as:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} (d_{ij}) & \rho_i \neq \max\limits_j (\rho_j) \\ \max\limits_j (d_{ij}) & \rho_i = \max\limits_j (\rho_j) \end{cases} \tag{3}$$

According to Definition 2, discriminant distance on each feature vector is its distance from nearest feature vector with higher local density. Specially, discriminant distance on the feature vector with maximum local density is the maximum distance between it and other feature vectors. For the feature vectors with large local density, they may be in the same cluster or may be in the different clusters. If their discriminant distance is large, the probability that they are in different clusters is great too. For this reason, discriminant distance on each feature vector characterizes the separability of different clusters in the dataset.

Definition 3 The decision graph of feature set X is a scatter plot with (ρ_i, δ_i), in which ρ_i is the horizontal ordinate that represents local density of feature vector f_i, and δ_i is the vertical ordinate that represents discriminant distance of feature vector f_i.

According to Definition 3, the feature points on the top-right corner of decision graph have larger local densities and discriminant distances, and they have higher probability to be clusters of features. Therefore, it is intuitive to evaluate feature importance with decision graph. For example, given two different means and covariance matrices, 40 points are generated randomly in two-dimensional plane, which contains two clusters, each of which has 20 points distributed normally and numbered with different colors in Fig. 1. From Fig. 1, we can intuitively find that the points with small distances have large similarities. The decision graph of points in Fig. 1 is shown in Fig. 2, in which point 7 and point 26 are obviously separated with others, and both of them may be cluster centers. In fact, it is identical with real case. Compared with Fig. 1, point 7 is the cluster center of first 20 points marked with red color, and point 26 is the cluster center of latter 20 points marked with blue color. Therefore, it is reasonable to identify cluster centers from decision graph.

2.3 Feature importance criteria

Firstly, suppose that feature set has typical cluster structure, i.e., the feature vectors with same or similar abilities of description should be clustered together. Then, each feature vector located in cluster center can be viewed as the most representative feature of the cluster. Since the distance between feature vectors with lower correlations or different abilities of description should be large, if the clusters of feature set are recognized, then the subset constructed by cluster centers can characterize the discriminant ability of original feature set sufficiently.

Based on the above idea, the feature importance evaluation criterion which is called decision graph score is presented as follows.

Definition 4 Decision graph score γ_i on feature vector f_i is defined as follows:

$$\gamma_i = \rho_i \cdot \delta_i \tag{4}$$

According to Definition 4, the larger the local density and discriminant of a feature vector are, the higher the decision graph score is, which corresponding to top-right

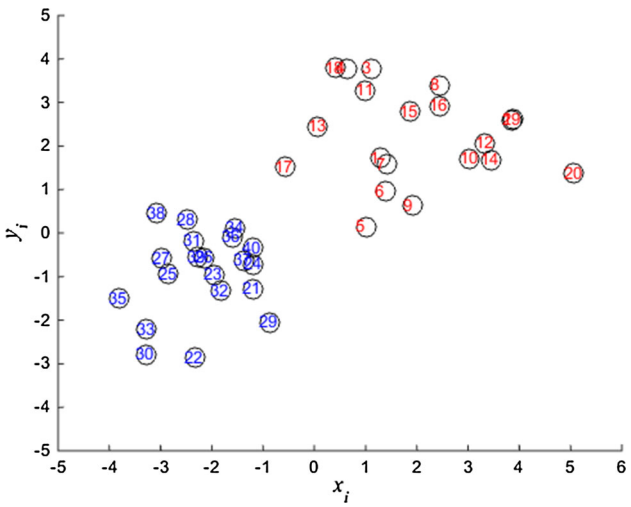


Fig. 1 Sample points in two-dimensional plane

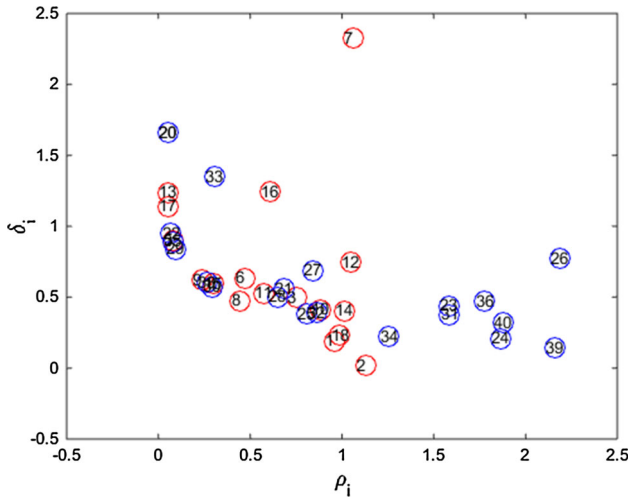


Fig. 2 Examples for decision graph

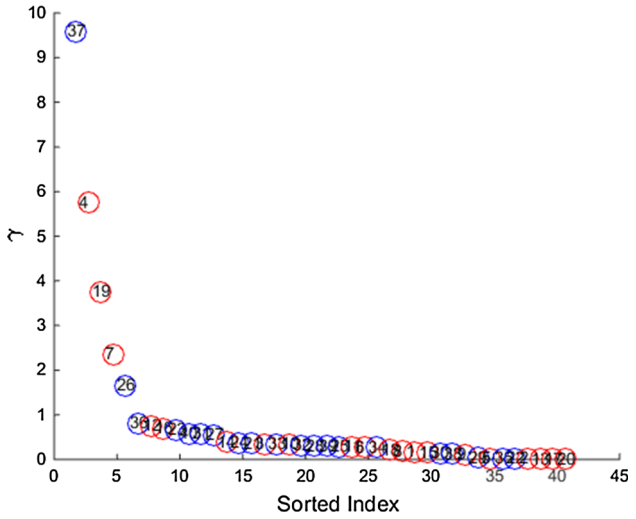


Fig. 3 Examples for decision graph Score

Table 1 Face datasets description

Dataset	Dimensionality	Number of samples	Number of classes
ORL	4096	400	40
YaleB	1024	2414	38
Altkom	2576	1200	80
PIE	1024	11,554	68
AR	1260	1400	100
MPEG7	2576	3175	635

area of decision graph and the features in such area are more important. By computing decision graph score of each feature vector, the features corresponding to first r largest scores can be selected as optimal feature subset, in which the correlation between features is smaller and can

Table 2 Experimental result comparisons on face datasets

Dataset	DGFS	Laplacian Score	MCFS	UDFS	NDFS	Relief-F	mRMR
ORL	**86.00 ± 2.98(45)**	67.75 ± 2.98(99)	**88.50 ± 2.71(94)**	**88.75 ± 5.30(64)**	67.75 ± 5.18(95)	80.50 ± 3.81(98)	48.75 ± 1.98(95)
YaleB	**78.25 ± 1.79(94)**	32.44 ± 1.56(100)	76.93 ± 0.88(100)	71.70 ± 2.59(98)	69.56 ± 1.55(97)	**81.89 ± 3.33(100)**	58.64 ± 4.54(99)
Altkom	**43.58 ± 4.70(99)**	32.25 ± 3.26(51)	21.75 ± 2.07(100)	27.67 ± 1.85(93)	20.92 ± 4.92(100)	25.92 ± 5.16(94)	15.58 ± 1.76(100)
PIE	94.30 ± 0.57(99)	82.21 ± 0.59(100)	94.83 ± 0.78(100)	**95.21 ± 0.58(96)**	94.88 ± 0.22(100)	94.03 ± 1.32(100)	83.59 ± 0.87(100)
AR	**72.00 ± 2.37(100)**	43.78 ± 3.26(100)	69.68 ± 2.84(98)	**72.75 ± 3.50(100)**	61.86 ± 6.16(97)	67.04 ± 3.78(100)	53.77 ± 2.91(100)
MPEG7	**59.34 ± 1.47(87)**	22.11 ± 1.87(100)	46.96 ± 0.21(100)	52.03 ± 1.63(99)	44.66 ± 1.94(100)	50.80 ± 5.17(100)	20.85 ± 0.99(99)

The comparison results with best performance are marked in bold

Table 3 Statistical test results on face datasets

Dataset	Laplacian Score	MCFS	UDFS	NDFS	Relief-F	mRMR
ORL	+(0.015733)	=(1)	=(0.30456)	+(2.7544e−05)	+(0.01423)	+(0.0002636)
YaleB	+(8.495e−07)	=(0.18385)	+(0.0093948)	+(0.0047223)	=(0.1772)	+(2.9372e−05)
Altkom	+(0.0052011)	+(4.6641e−06)	+(0.00035143)	+(8.029e−05)	+(0.016538)	+(7.7615e−05)
PIE	+(1.8048e−05)	=(0.611521)	−(0.033101)	=(0.589661)	=(0.66634)	+(2.657e−05)
AR	+(0.00013818)	=(0.86053)	=(0.27486)	+(0.0052046)	+(0.03158)	+(6.056e−05)
MPEG7	+(1.8105e−05)	+(0.0031185)	+(0.0024205)	+(1.9617e−06)	+(0.00055643)	+(7.3787e−06)
+/=/−	6/0/0	2/4/0	3/2/1	5/1/0	4/2/0	6/0/0

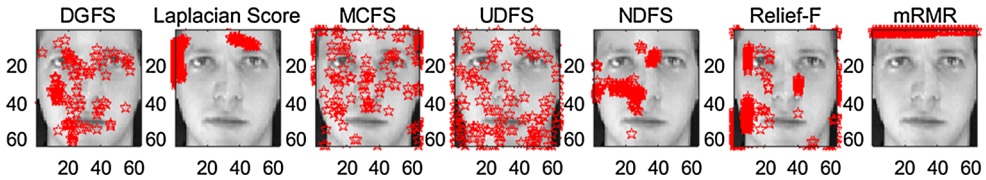


Fig. 4 Feature selection on ORL dataset

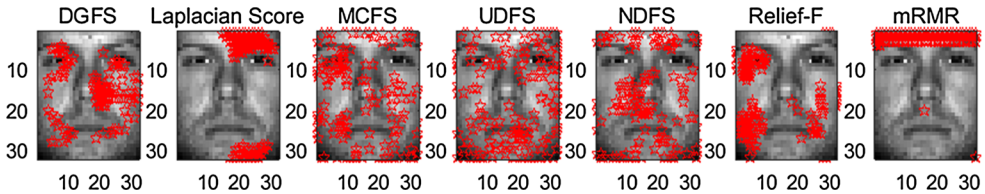


Fig. 5 Feature selection on YaleB dataset

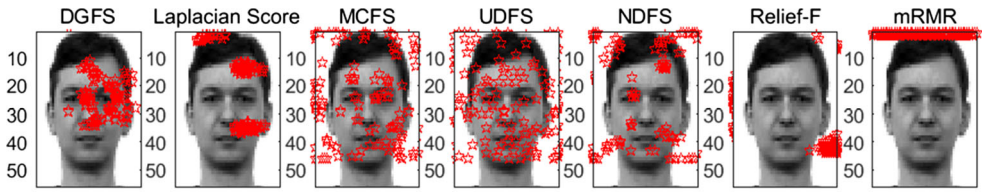


Fig. 6 Feature selection on Altkom dataset

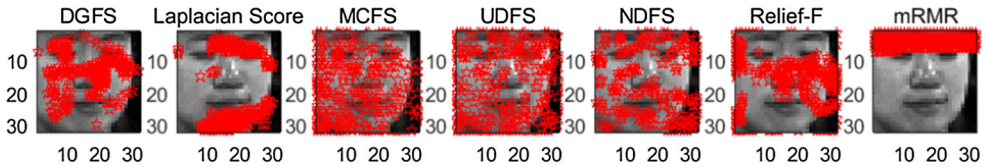


Fig. 7 Feature selection on PIE dataset

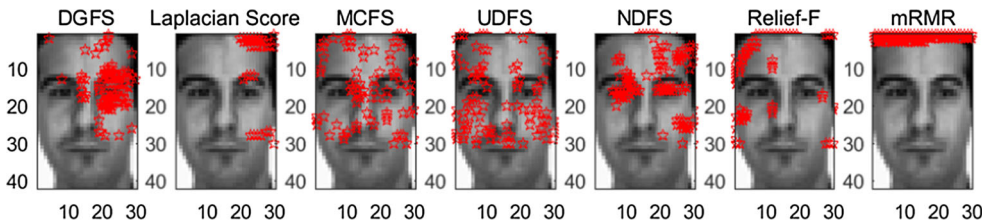


Fig. 8 Feature selection on AR dataset

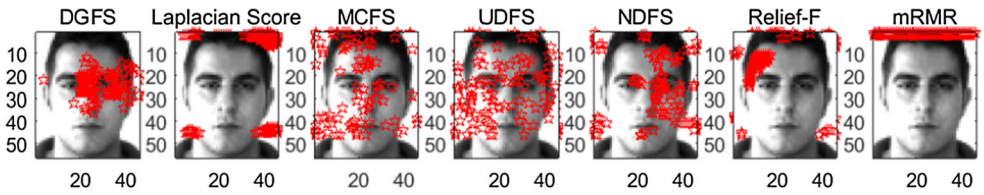
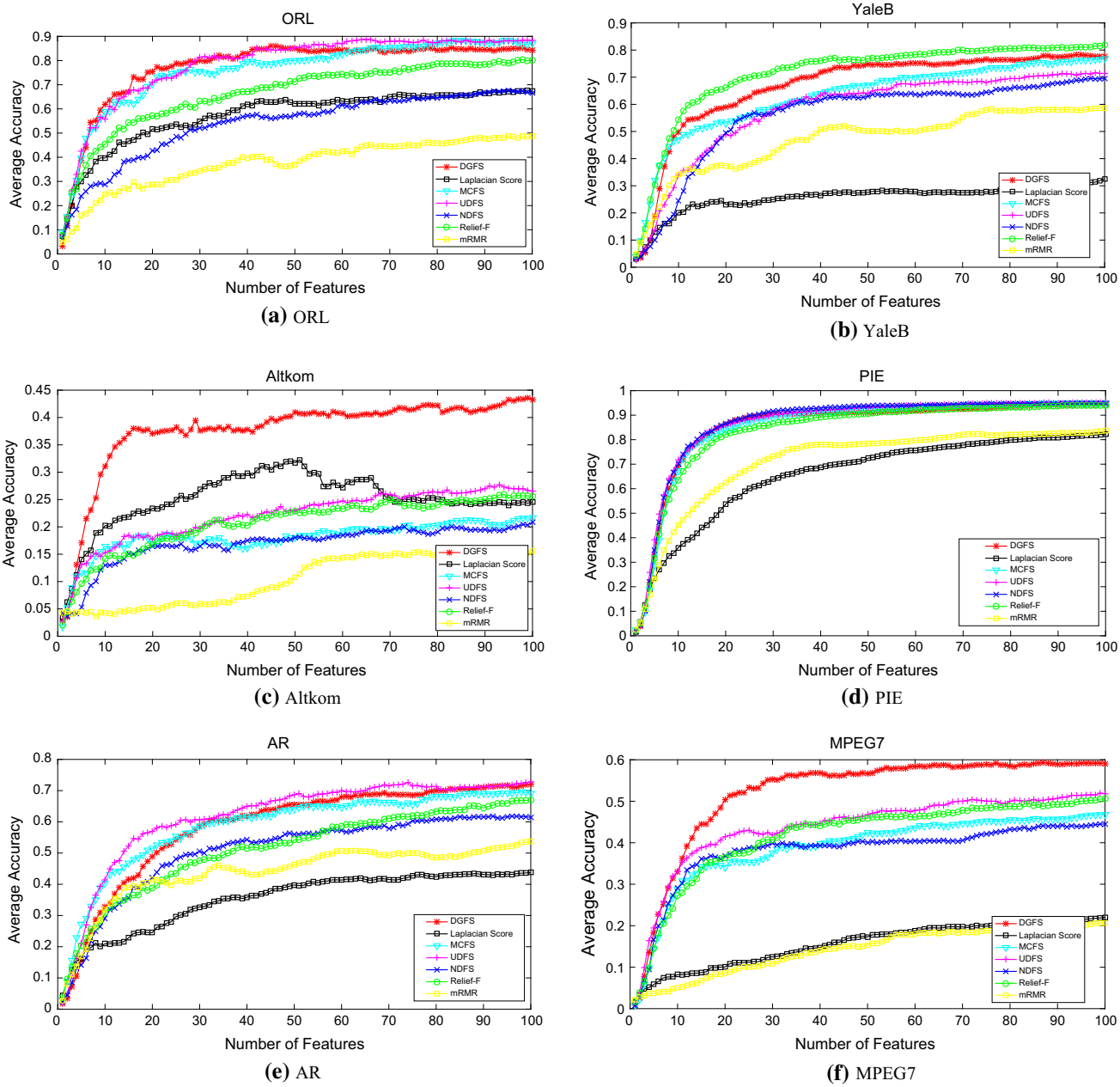


Fig. 9 Feature selection on MPEG7 dataset



(a) ORL **(b)** YaleB **(c)** Altkom **(d)** PIE **(e)** AR **(f)** MPEG7

Fig. 10 Average accuracy versus different numbers of selected features on face dataset

be selected as better representatives that contain information of original feature subset. Figure 3 shows decision graph scores of 40 sample points in Fig. 1 with descending order, where the first two sample points with highest decision graph scores are No. 37 and No. 4 that deviates from real case, but very close to real cluster center.

Based on the above definition and analysis, the procedure of DGFS algorithm is summarized as follows:

Table 4 Computational time costs comparisons on face datasets (*s*)

Dataset	DGFS	Laplacian Score	MCFS	UDFS	NDFS	Relief-F	mRMR
ORL	3.20	**0.08**	1.05	467.61	78.71	0.59	2.86
YaleB	0.38	**0.37**	1.85	30.17	12.61	0.95	5.35
Altkom	1.63	**0.27**	2.58	121.56	32.56	1.26	3.92
PIE	**1.82**	6.40	12.30	483.05	170.79	4.73	18.39
AR	0.40	**0.18**	1.24	22.52	9.04	0.65	4.13
MPEG7	3.49	**1.11**	6.64	202.41	64.43	4.08	8.26

The comparison results with best performance are marked in bold

Table 5 UCI datasets description

Dataset	Dimensionality	Number of samples	Number of classes
Heart	13	296	2
Vote	16	435	2
Dermatology	34	366	6
Australian	14	690	2
Wine	13	178	3
Credit	15	690	2
Car	6	1728	4
E. coli	8	336	8
Seeds	7	210	3
WDBC	30	569	2

Input: feature matrix X, number of selected features r;

Output: feature subset Y.

Procedure:

Step 1: Compute local densities and discriminant distances of feature matrix X according to Eqs. (1) and (3).

Step 2: Compute decision graph score $\gamma_i = \rho_i \cdot \delta_i$ according to Eq. (4).

Step 3: Sort decision graph scores with descending order, and select features with first r largest scores as optimal feature set Y.

3 Experimental results

3.1 Experimental setting

In order to verify the effectiveness of the proposed method, in the experiments, we compare DGFS with other unsupervised feature selection methods, such as Laplacian score, multi-cluster feature selection (MCFS), unsupervised discriminant feature selection (UDFS) and nonnegative discriminant feature selection (NDFS). Before feature selection, data deduplication is performed on the rows and columns of data matrix, and then, all feature values are normalized into vectors with unit norm. A good feature selection method should make the selected feature subset get a better classification result even by the simple classifier, such as k nearest neighbor. In the experiments, fivefold cross-validation is used to evaluate each method. Firstly, the original high-dimensional dataset is divided into five parts randomly, four parts of which as the training set and one part of which as a test set in turn. Then, feature selection methods are conducted on training set in each partition, and the indices of selected features can also be used in feature selection on testing set. Finally, the average classification accuracy by nearest neighbor classifier on selected features is reported. For the purpose of exploring the statistical significance of the results, we performed t test to statistically compare methods on multiple datasets.

In the experiments, parameters of each algorithm are empirically set as follows: In DGFS, local density of each sample is computed by Eq. (2), where the truncated distance d_c is set as the distance at the position of two percent of total distances between feature vectors with ascending order. In Laplacian Score, neighborhood parameter k is set to 5, and similarity between feature vectors is computed by cosine metric. In MCFS, neighborhood parameter k is set to 5; in UDFS, regularization parameter is set to 0.01; and in NDFS, parameter neighborhood is set to 5, and similarity between feature vectors is computed by cosine metric, and the number of maximum iterative steps is set to 30, and regularization is set to 0.1.

3.2 Classification results on face datasets

Grayscale image dataset is typically high-dimensional after stacking columns into a vector, which contains large amount of redundant, irrelevant and noisy pixels, so it is

Table 6 Experimental result comparisons on UCI datasets

Dataset	DGFS	Laplacian Score	MCFS	UDFS	NDFS	Relief-F	mRMR
Heart	**79.47 ± 7.18(7)**	65.02 ± 5.61(13)	69.99 ± 8.20(11)	70.33 ± 7.94(7)	65.36 ± 5.86(11)	71.68 ± 6.93(5)	65.02 ± 5.61(13)
Vote	**93.77 ± 4.25(7)**	**93.32 ± 2.26(9)**	**93.33 ± 2.09(12)**	90.55 ± 4.01(16)	91.03 ± 4.71(12)	**94.25 ± 2.44(3)**	90.32 ± 3.94(16)
Dermatology	**95.10 ± 2.01(33)**	**94.54 ± 1.90(34)**	**95.10 ± 2.01(33)**	**94.54 ± 1.90(34)**	**94.81 ± 1.78(25)**	**94.54 ± 2.16(30)**	**95.35 ± 1.59(31)**
Australian	**80.43 ± 1.99(6)**	67.23 ± 5.65(13)	**81.01 ± 2.49(4)**	68.28 ± 8.22(4)	67.08 ± 6.14(14)	75.95 ± 7.74(4)	68.10 ± 4.79(12)
Wine	**95.51 ± 3.79(5)**	84.87 ± 4.52(13)	87.03 ± 7.12(4)	85.41 ± 4.50(10)	84.87 ± 4.52(13)	84.87 ± 4.52(13)	84.87 ± 4.52(13)
Credit	**82.61 ± 3.03(10)**	75.08 ± 5.09(6)	77.82 ± 3.59(13)	76.52 ± 3.18(14)	**82.90 ± 4.67(7)**	75.07 ± 3.23(15)	76.38 ± 3.60(13)
Car	**90.10 ± 1.47(5)**	84.61 ± 1.88(6)	**88.60 ± 4.48(5)**	84.55 ± 1.48(6)	84.95 ± 2.01(6)	84.49 ± 1.74(6)	84.78 ± 2.04(6)
E. coli	**97.91 ± 1.72(1)**	95.53 ± 2.76(7)	95.53 ± 2.76(7)	95.53 ± 2.76(8)	95.53 ± 2.76(8)	95.53 ± 2.76(8)	95.53 ± 2.76(8)
Seeds	**95.71 ± 3.10(4)**	92.38 ± 4.26(6)	92.86 ± 4.76(5)	92.38 ± 4.26(7)	92.38 ± 4.26(7)	92.38 ± 4.26(7)	92.38 ± 4.26(7)
WDBC	**94.02 ± 1.59(20)**	91.04 ± 2.44(19)	92.98 ± 2.45(4)	91.04 ± 2.44(30)	91.04 ± 2.44(9)	91.04 ± 2.73(17)	92.09 ± 2.50(21)
+/=/−	–	8/2/0	5/5/0	9/1/0	8/2/0	8/2/0	9/1/0

The comparison results with best performance are marked in bold

necessary to select features before image analysis. The classification results on six benchmark face image datasets, such as ORL, YaleB, Altkom, PIE, AR and MPEG7, are reported in the section.

Face datasets used in the experiments are given in Table 1, wherein the ORL[1] dataset is created by Bell Labs of the University of Cambridge, which contains 400 images, including 40 individual facial expressions (open eyes or closed eyes, smiling or not), occlusion (wearing glasses or not) and slight changes of pose; YaleB[2] dataset is created by computer vision and control center in Yale university, which contains 38 individuals under strictly controlled conditions of illumination and poses; Altkom[3] dataset contains 80 individuals with 15 images for each individual; PIE[4] dataset is created by Carnegie Mellon University, which contains 41,368 face images of 68 individuals under strictly controlled conditions of pose, illumination and expression. The AR dataset[5] contains more than 4000 frontal images from 126 persons (70 men and 56 women) with different facial expressions, lighting conditions and occlusions. In the experiment, we choose a subset which contains 50 males and 50 females. For each person, 14 images with only illumination and expression changes are selected. MPEG-7 content set of face images[6] was provided by the Heinrich Hertz Institute of Germany, which contains 3175 face images of 635 persons.

The classification results of each algorithm on different datasets are given in Table 2, in which the average classification accuracy (%), standard deviation (%) and the corresponding number of features that achieved the highest average classification accuracy on five cross-validation experiments are reported. From Table 2, we can see that the average classification accuracy of Laplacian Score is much lower than the other three methods, while in most cases, the proposed DGFS method achieves the higher average classification accuracy. In order to identify the pairwise different significance between the proposed DGFS method and other compared methods, *t* test is used to make decisions for the null hypothesis that the pairwise difference of optimal average accuracies between two methods has a mean equal to zero. The *t* test results at the 5% significance level on the six face datasets are reported in Table 3, where symbol '+' denotes the DGFS method that outperforms the compared method significantly according to t test, while '−' denotes the compared method

[1] http://www.uk.research.att.com/facedatabase.html.

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html.

[3] http://www.iis.ee.ic.ac.uk/icvl/code.htm.

[4] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html.

[5] http://rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html.

[6] http://www.darmstadt.gmd.de/mobile/hm/projects/MPEG7/Documents/N2466.html.

Table 7 Statistical test results on UCI datasets

Dataset	Laplacian Score	MCFS	UDFS	NDFS	Relief-F	mRMR
Heart	+(0.033019)	+(0.015613)	+(0.016983)	+(0.00418)	+(0.02825)	+(0.024059)
Vote	=(0.43256)	=(0.13041)	+(0.031872)	+(0.038345)	=(0.2483)	+(0.01968)
Dermatology	=(0.17782)	=(0.58699)	=(0.38322)	=(0.6313)	=(0.99355)	=(0.98877)
Australian	+(0.031481)	=(0.34256)	+(0.028825)	+(0.032562)	+(0.04916)	+(0.028825)
Wine	+(0.033018)	+(0.042956)	+(0.029358)	+(0.033018)	+(0.033018)	+(0.033018)
Credit	+(0.0034361)	+(0.011273)	+(0.005262)	=(0.18827)	+(0.026645)	+(0.049611)
Car	+(0.00019646)	=(0.3739)	+(0.00066934)	+(0.00025654)	+(0.00043762)	+(0.00025788)
E. coli	=(0.24371)	=(0.24371)	=(0.24371)	=(0.24371)	=(0.24371)	=(0.24371)
Seeds	+(0.04057)	+(0.04657)	+(0.04057)	+(0.04057)	+(0.04057)	+(0.04057)
WDBC	+(0.006312)	+(0.006312)	+(0.04909)	+(0.042295)	+(0.048797)	+(0.029083)
+/=/−	7/3/0	5/5/0	8/2/0	7/3/0	7/3/0	8/2/0

that outperforms the DGFS method, and '=' denotes that there is no statistically significant difference between the results obtained by the DGFS method and the compared method. The obtained p values in t test are reported in parentheses. Table 3 shows that the proposed DGFS method has achieved a statistically higher accuracy than all other compared methods on all the six face datasets in most cases. Compared with Laplacian Score and mRMR, DGFS is always better, while there is only one case that MCFS, UDFS and Relief-F perform better than DGFS.

In order to analyze physical meaning of the selected feature by each method intuitively, the optimal selected features marked with red stars on the six face datasets are shown in Figs. 4, 5, 6, 7, 8 and 9, in which the selected features by Laplacian Score locate in continuous local area of face image, and the pixels concentrated in those locations have small changes and with typical manifold structure, but cannot distinguish between different faces. While the selected pixel features by MCFS, UDFS and NDFS are more dispersed, the selected pixels by DGFS are mostly concentrated on the positions of eyes, nose and mouths, which also show that these parts play a key role for distinguishing different facial images, which is also consistent with knowledge of human cognitive experience.

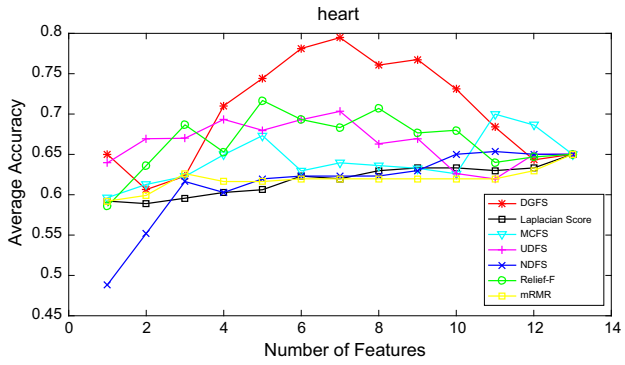
In addition, Fig. 10 shows the changes of average classification accuracy with the number of selected features each face dataset changes. In the experiment, the maximum number of features we set as 100. As shown in Fig. 10, with the increasing number of features, the classification accuracy is also increasing. Within the 20 features, the classification accuracy increases fast; after that, with the number of features increasing, the accuracy increases slowly. When they achieve the best results, the accuracy of the most methods begins to stay stable. From Table 2 and Fig. 10, we can see that, in most cases, the proposed DGFS method not only selects a much smaller numbers of features, but results in better classification performance as well.

Table 4 lists the CPU time in seconds obtained from the different algorithms on the six datasets. The UDFS and NDFS work the poorest in terms of CPU time, while the computational time costs of other methods are comparable. This is due to the fact that the iterative optimizing processes in UDFS and NDFS models are time-consuming when the number of dimensionality of samples is large.

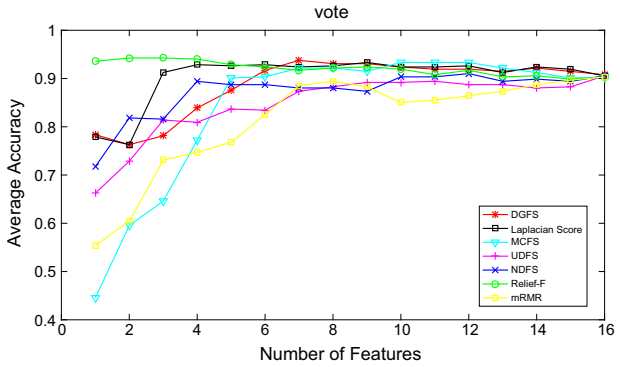
3.3 Classification results on UCI datasets

To further compare the effect of each feature selection algorithm, ten most commonly used UCI datasets[7] that are from real-world applications, such as health, political and economic fields, are used in the experiments, which are given in Table 5, where Heart dataset is the diagnosis data of patients with heart disease, Vote dataset is the voting data from Republican and Democratic Congress, Dermatology dataset is the clinical and histopathological data for six kinds of skin diseases, Australia dataset is the Australia's credit approval data, Wine dataset is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars, Credit dataset concerns credit card applications, Car dataset is the car evaluation database, *E. coli* dataset contains protein localization sites information, Seeds dataset is the measurement of geometrical properties of kernels belonging to three different varieties of wheat, and WDBC dataset is the diagnostic Wisconsin breast cancer database. In the preprocessing step, the missing values in original datasets are manually set to 0 and non-numeric category attributes are represented as integers.

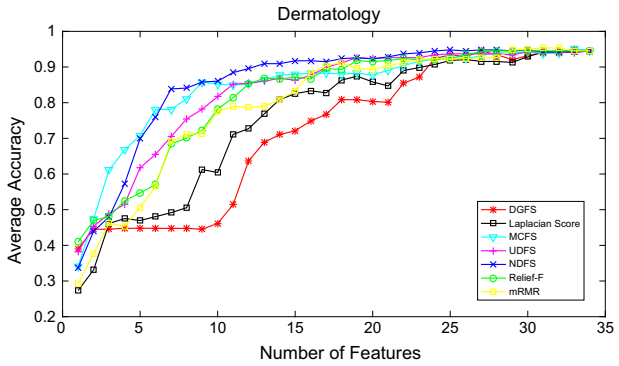
[7] http://archive.ics.uci.edu/ml/datasets.html.



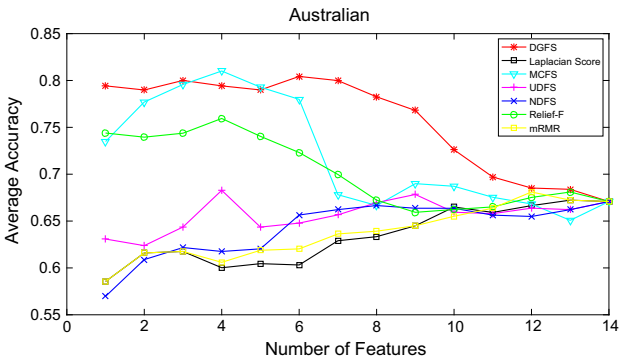
(a) Heart



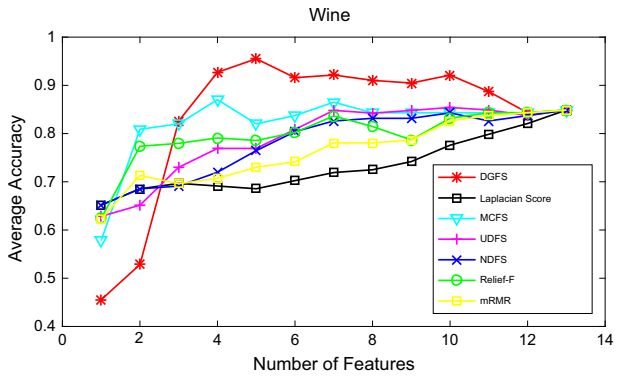
(b) Vote



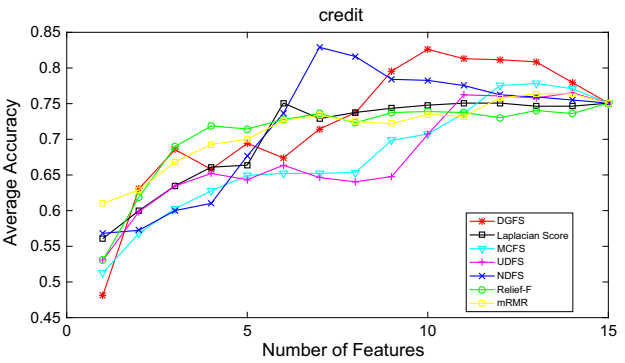
(c) Dermatology



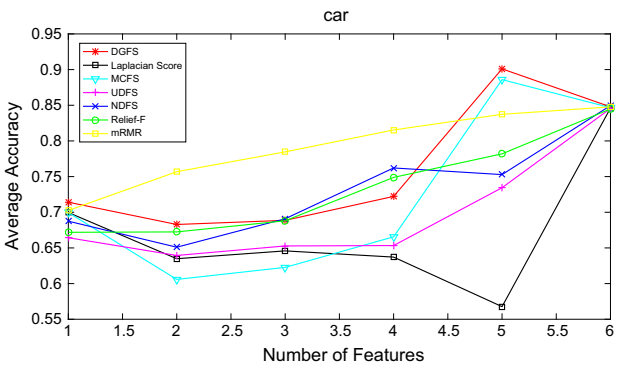
(d) Australian



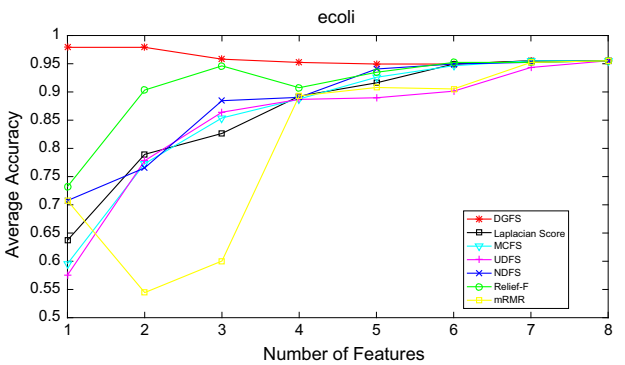
(e) Wine



(f) Credit

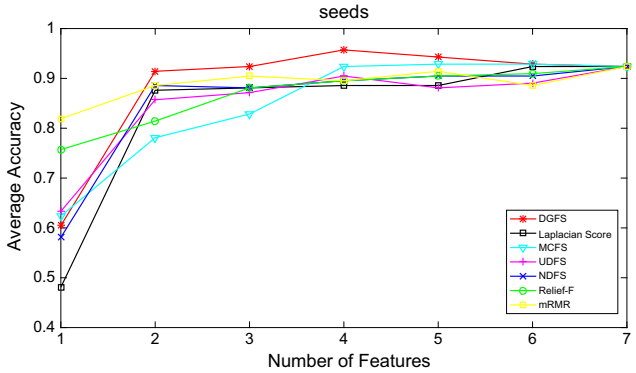


(g) Car

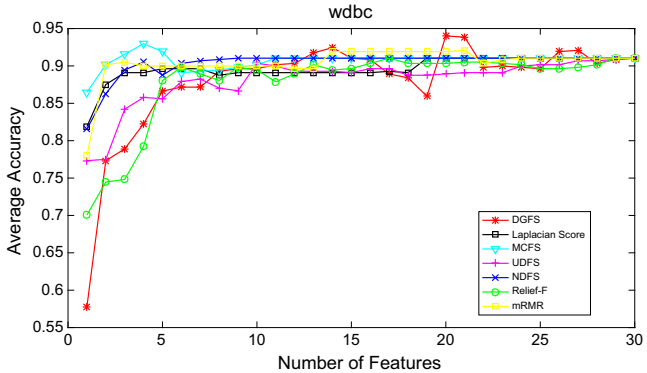


(h) Ecoli

Fig. 11 Average accuracy versus different numbers of selected features on UCI dataset



(i) Seeds



(j) WDBC

Fig. 11 continued

The classification results of each algorithm on different datasets are given in Table 6, in which the average classification accuracy (%), standard deviation (%) and the corresponding number of selected features that achieved the highest average classification accuracy on five cross-validation experiments are reported. The significance of average accuracy differences is identified by the t test, in which the null hypothesis is that the pairwise difference of optimal average accuracies between any two methods comes from a normal distribution with mean equal to zero and unknown variance at the 5% significance level. As shown in Table 7, DGFS performs better than other methods.

Figure 11 shows the average classification accuracies with different number of selected features. As shown in Fig. 11, in Heart and Wine datasets, DGFS performs significantly better than other methods, and the high accuracy rate has been achieved with less feature values. While in Vote, Dermatology and Seeds datasets, with the increase of the number of features, the accuracy values of different methods are tend to be the same. In Australian dataset, the impact of the number of features on the results is quite unstable. For instance, upon analyzing the results of DGFS, MCFS and Relief-F methods, classification performances tend to decrease despite the increasing feature values. For Credit dataset, high accuracy rate has been achieved with DGFS method. A value close to this rate of accuracy has been obtained by using 7 features with NDFS method. In Car dataset, DGFS and MCFS have comparable performances and perform better than others. High accuracy value has been achieved with the proposed method by using only one feature in *E. coli* dataset. In WDBC dataset, high accuracy rate has been obtained by using 20 features with DGFS method, while other methods perform comparably with the increasing number of features.

4 Discussion and conclusion

In this paper, based on the concept of decision graph, an unsupervised feature selection method is proposed. Since the process of feature selection can be viewed as feature clustering, the features as cluster centers can not only be representative features of other features in the same cluster, but also can discriminate with features in other clusters. Therefore, the selected features have less redundancy and can preserve the inherent information contained in original feature set. To identify cluster centers of feature set, we introduced the definitions of local density and discriminant distance, which can be used to construct decision graph for identifying cluster centers. Then, for evaluating feature importance, the index named decision graph score is proposed. Feature selection can be achieved by decision graph score ranking.

In the experiments, the performance of DGFS method is evaluated on 16 publicly available real-life datasets, including 6 face datasets and 10 UCI datasets. The number of features for these datasets varies from 6 to 4096, and the number of samples ranges from 178 to 11,554. From Tables 3 and 7, we can see that, at the significance level of 0.05, the proposed DGFS is statistically superior than the state-of-the-art feature selection methods for data classification, regardless of dimensionality and distributional shape of data samples.

Acknowledgements We thank the editors and anonymous reviewers for their very useful comments and suggestions. This work is supported in part by Science Computing and Intelligent Information Processing of Guangxi Higher Education Key Laboratory (No. GXSCIIP201406), Doctoral Starting up Foundation of Northwest A&F University (No. 2452015302), the Fundamental Research Funds for the Central Universities (No. 2452015197), National Natural Science Foundation of China (No. 61402481) and Hebei Province Natural Science Foundation of China (No. F2015403046).

Compliance with ethical standards

Conflict of interest The authors declared that they have no conflicts of interest to this work. All authors of this manuscript have directly participated in planning, execution and analysis of this study. The contents of this manuscript have not been copyrighted or published previously, or under consideration for publication elsewhere.

References

1. Chen B, Hong J, Wang Y (1997) The problem of finding optimal subset of features[J]. Chin J Comput 20(2):133–138
2. Kira K, Rendell LA (1992) The feature selection problem: traditional methods and a new algorithm[C]. In: Proceedings of the ninth national conference on artificial intelligence, Menlo Park, CA, USA, pp 129–134
3. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF[J]. Mach Learn 53(1–2):23–69
4. Almuallim H, Dietterich TG (1994) Learning boolean concepts in the presence of many irrelevant features[J]. Artif Intell 69(1):279–305
5. Hall MA (1999) Correlation-based feature selection for machine learning[D]. The University of Waikato, Hillcrest
6. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution[C]. ICML 3:856–863
7. Nie F, Xiang S, Jia Y, Zhang C, Yan S (2008) Trace ratio criterion for feature selection[C]. In: Proceedings of conference on artificial intelligence (AAAI), pp 671–676
8. Wei HL, Billings SA (2007) Feature subset selection and ranking for data dimensionality reduction[J]. IEEE Trans Pattern Anal Mach Learn 29(1):162–166
9. Li WH, Chen WM, Yang LP, Gong WG (2007) Face feature selection and recognition based on different types of margin[J]. J Electron Tech 29(7):1744–1748
10. Dash M, Liu H (2003) Consistency-based search in feature selection[J]. Artif Intell 151(1):155–176
11. Dash M, Choi K, Scheuermann P et al (2002) Feature selection for clustering-A filter solution [C]. In: Proceedings of the 2nd IEEE international conference data mining, Piscataway, pp 115–122
12. Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity[J]. IEEE Trans Pattern Anal Mach Intell 24(3):301–312
13. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. Pattern Anal Mach Intell IEEE Trans 27(8):1226–1238
14. Singh P (2014) Devanagri handwritten numeral recognition using feature selection approach[J]. Int J Intell Syst Appl 12:40–47. doi:10.5815/ijisa.2014.12.06
15. Junling X, Yuming Z, Lin C, Baowen X (2012) An unsupervised feature selection approach based on mutual information[J]. J Comput Res Dev 49(2):372–382
16. Bandhyapadhyay S, Bhadra T, Mitra P, Maullik U (2014) Integration of dense subgraph finding with feature clustering for unsupervised feature selection[J]. Pattern Recogn Lett 40:104–112
17. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection[C]. In: Advances in neural information processing systems, pp 507–514
18. Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning[C]. In: Proceedings of the international conference on machine learning (ICML), Corvallis, Oregon, pp 1151–1157
19. Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data[C]. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 333–342
20. Yang Y, Shen H T, Ma Z, et al (2011) l2, 1-norm regularized discriminative feature selection for unsupervised learning[C]. In: IJCAI proceedings-international joint conference on artificial intelligence, vol 22, no 1, p 1589
21. Li Z, Yang Y, Liu J et al (2012) Unsupervised feature selection using nonnegative spectral analysis[C]. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence, Canada, pp 1026–1032
22. Du L, Shen Z, Li X et al (2013) Local and global discriminative learning for unsupervised feature selection[C]. Data mining (ICDM), 2013 IEEE 13th international conference on IEEE, pp 131–140
23. Zhao Z, Wang L, Liu H et al (2013) On similarity preserving feature selection[J]. Knowl Data Eng IEEE Trans 25(3):619–632
24. Liu Tao W, Gongyi CZ (2005) An effective unsupervised feature selection method for text clustering[J]. J Comput Res Dev 42(3):381–386
25. Lu Y, Cohen I, Zhou X S et al (2007) Feature selection using principal feature analysis[C]. In: Proceedings of the 15th international conference on Multimedia. ACM, pp 301–304
26. Song Q, Ni J, Wang G (2013) A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. Knowl Data Eng IEEE Trans 25(1):1–14
27. Yan H, Yang J (2015) Sparse discriminative feature selection[J]. Pattern Recogn 48(5):1827–1835
28. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks[J]. Science 344(6191):1492–1496
29. Brahim AB, Limam M (2016) A hybrid feature selection method based on instance learning and cooperative subset search[J]. Pattern Recogn Lett 69:28–34
30. Şen B, Peker M, Çavuşoğlu A et al (2014) A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms[J]. J Med Syst 38(3):1–21