Contents lists available at ScienceDirect

Neurocomputing

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



CrossMark

Feature selection with effective distance

Mingxia Liu a,b, Daoqiang Zhang a,*

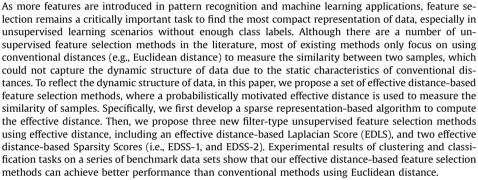
- a School of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China
- ^b School of Information Science and Technology, Taishan University, Taian 271021, China



Article history: Received 26 January 2015 Received in revised form 18 July 2015 Accepted 19 July 2015 Available online 8 June 2016

ARTICLE INFO

Keywords: Feature selection Effective distance Classification Clustering Sparse representation



© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In many pattern recognition and machine learning applications, the number of features (or variables) is becoming much higher, and is even higher than that of the observations [1-6]. For example, there are usually tens of thousands of features in neuroimaging data, and the number of subjects is very limited, which is called the small-sample-size problem [7]. In the literature, feature selection (or variable selection) has been shown effective in solving the small-sample-size problem by reducing feature dimension to eliminate noisy or redundant features, and, thus, can help improve learning performances and facilitate data understanding [8-14].

According to the different using of class labels, existing feature selection methods can be divided into two groups [15-18], i.e., supervised feature selection methods and unsupervised ones. Specifically, supervised feature selection methods require full class labels for training samples, while it is usually difficult or costly to obtain enough class labels, especially in the case with many categories [19-23]. In contrast, unsupervised feature selection methods that do not need any class label information have received much attention in recent years [12,24-26]. In general, most

* Corresponding author. E-mail address: dqzhang@nuaa.edu.cn (D. Zhang). of existing unsupervised feature selection methods use conventional distances (e.g., Euclidean distance) to measure the similarity between two samples. For example, Gao et al. [27] proposed a hypergraph-based 3D object retrieval method and achieved stateof-the-art results, by using Euclidean distance as the similarity measure. However, since conventional distances only focus on the magnitude of similarity between a pair of samples, they cannot reflect the dynamic structure information of original data, which may lead to sub-optimal performance for subsequent learning

Recently, a probabilistically motivated distance measure called effective distance is proposed to reflect the dynamic structure of data [28], with an illustration shown in Fig. 1. In Fig. 1(a), there is a graph with three nodes with equal weight quantified by gray arrows. In Fig. 1(b), we compute the probability P(n|m) of a random walker at node m of moving to n, where the magnitude of this probability is indicated by the line width. For example, the probability P(B|A) of a random walker at node A of moving to node B is defined as 1/6, where the number 6 denotes the connection number of node A to the other nodes. From Fig. 1(b), one can see that a random walker starting at A has a smaller probability (i.e., P (B|A)=1/6) of going to B than vice versa, while that starting at C has a larger probability (i.e., P(B|C)=1) of going to B than vice versa. In the definition of effective distance [28], a small probability P(n|m) denotes that the effective distance from note m to node n is long, and vice versa. Compared to conventional

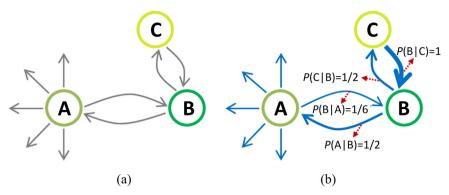


Fig. 1. Graphical explanation of effective distance. (a) Consider a graph with three nodes with equal edge weight quantified by gray arrows. (b) We can compute the probability P(n|m) of a random walker at node m of moving to n. Note that the magnitude of these probabilities is indicated by the line width. For example, a random walker starting at A has a smaller probability of going to B than vice versa, while that starting at C has a larger probability of going to B than vice versa.

geographic distance, effective distance helps in revealing hidden pattern geometry of data, by considering underlying dynamic structure information of data [28]. Intuitively, replacing conventional Euclidean distance with such effective distance can bring dynamic information to learning tasks, and thus, may help promote the learning performance. To the best of our knowledge, no previous work has adopted such effective distance as similarity measure in feature selection domain.

Accordingly, in this paper, we propose a set of effective distance-based feature selection methods, by replacing conventional distance with effective distance. To be specific, we first develop a sparse representation-based algorithm to construct a bi-directional network, in order to obtain the effective distance. Then, we develop three novel unsupervised filter-type feature selection methods by using such effective distance, including an effective distance-based Laplacian Score (EDLS), and two effective distancebased Sparsity Scores (i.e., EDSS-1, and EDSS-2). Experimental results on a series of benchmark data sets demonstrate the efficacy of our methods. The major contributions of this paper are twofold. First, we propose a novel effective distance computation method by using sparse representation, which can take advantage of the global property of data. Second, we develop three unsupervised feature selection methods (i.e., EDLS, EDSS-1 and EDSS-2) by using effective distance as the similarity measure.

The rest of this paper is organized as follows. We first briefly introduce some background information on feature selection and the effective distance in Section 2. In Section 3, we propose a sparse representation-based algorithm for computing the effective distance, and then develop three new filter-type feature selection methods by using such effective distance. Afterwards, we elaborate the experiments and comparative results in Section 4. Finally, we draw conclusion and point out future research directions in Section 5.

2. Related work

2.1. Feature selection

In general, existing feature selection methods can be roughly categorized into two groups, including 1) filter-type feature selection methods, and 2) wrapper-type feature selection methods [29]. Specifically, with a pre-defined learning algorithm, Wrapper-type methods aim to find an optimal feature subset by evaluating the performance of each candidate feature subset [30–33]. Hence, wrapper-type methods often achieve better performance in terms of accuracy than filter-type methods, but are usually computationally more expensive [34]. In contrast, filter-type methods involve no learning algorithm, and select features according to

mutual information, correlation, or other criteria [35–39]. In practice, filter-type methods are usually adopted due to their simplicity and computational efficiency, especially in the case with high-dimensional features [40,41]. Among various filter-type feature selection methods, feature ranking methods remain popular in machine learning and pattern recognition domains. Specifically, feature ranking methods consider features individually and achieve a ranked list of selected features ordered by their importance [42-45]. Thus, feature ranking methods are usually computationally more efficient and are very scalable to data sets with huge number of samples and high dimensionality [2,37,46]. In this work, we focus on filter-type feature selection methods. In the literature, Variance [47], Laplacian Score [48], Fisher Score [47]. Constraint Score [17] and Sparsity Score [49] are typical feature ranking algorithms. In the following, we briefly introduce several feature ranking methods that are highly related to our proposed methods.

Given a set of data samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$, where $\mathbf{x}_i \in \mathbb{R}^d$, N is the number of data points, and d is the feature dimension. Let f_{ri} denote the r-th feature of the i-th sample \mathbf{x}_i . Denote the mean of the r-th feature as $\mu_r = \frac{1}{N} \sum_{i=1}^N f_{ri}$. As a simple unsupervised method, Laplacian Score (LS) prefers features with larger variances as well as stronger locality preserving ability. A key assumption in Laplacian Score is that the data points from the same class should be close to each other. The Laplacian score of the r-th feature denoted as LS_r , which should be minimized, is computed as follows [48]:

$$LS_{r} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (f_{ri} - f_{rj})^{2} S_{ij}}{\sum_{j=1}^{N} (f_{ri} - \mu_{r})^{2} D_{ii}}$$
(1)

Here **D** is a diagonal matrix and $D_{ii} = \sum_{j=1}^{N} S_{ij}$, where S_{ij} is defined by the neighborhood relationship between samples \mathbf{x}_i and \mathbf{x}_j as follows:

$$S_{ij} = \begin{cases} e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\sigma}}, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases}$$
 (2)

where σ is a constant to be set.

As another unsupervised feature ranking algorithm, Sparsity Score is recently proposed, where a l_1 graph adjacency structure is used to guide the feature selection process. Give a training data set, an l_1 graph represented by a sparse reconstruction weight matrix \boldsymbol{S} is first constructed via sparse representation [49–51], where the element of \boldsymbol{S} define the similarity between any pair of samples. Then, based on the matrix \boldsymbol{S} , the Sparsity Score (denoted as SS-1) of the r-th feature (SS_r^1), which should be minimized, as follows:

$$SS_r^1 = \sum_{i=1}^{N} \left(f_{ri} - \sum_{j=1}^{N} \hat{\mathbf{s}}_{i,j} f_{rj} \right)^2 = \mathbf{f}_r^T (\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S} \mathbf{S}^T) \mathbf{f}_r$$
(3)

where $\hat{s}_{i,j}$ is the entry of the sparse reconstruction weight matrix S constructed using all data points. By minimizing SS_r^1 , features that can best respect the pre-defined l_1 graph structure are regarded as being more discriminative.

Then, by considering the variance information in a specific feature space, another Sparsity Score (denoted as SS-2) of the r-th feature (SS_r^2) is defined as

$$SS_r^2 = \frac{\sum_{i=1}^N \left(f_{ri} - \sum_{j=1}^N \hat{\mathbf{s}}_{i,j} f_{rj} \right)^2}{\frac{1}{N} \sum_{i=1}^N \left(f_{ri} - \mu_r \right)^2} = \frac{\mathbf{f}_r^T (\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S} \mathbf{S}^T) \mathbf{f}_r}{\mathbf{f}_r^T \left(\mathbf{I} - \frac{1}{N} \mathbf{1} \mathbf{1}^T \right) \mathbf{f}_r}$$
(4)

In Eqs. (3) and (4), we prefer features that can best preserve the l_1 graph structure and those with large variance that have stronger representative ability. That is, with smaller reconstruction error (i.e. to preserve the l_1 graph structure), as well as larger variance for r-th feature, Sparsity scores tend to be small that means the feature would be more important.

2.2. Effective distance

Many feature selection algorithms depend on calculating distances between samples, e.g., Laplacian Score (LS) [48], Constraint Score (CS) [17], and Sparsity Score (SS) [49]. Some recent works [52] further investigate Hausdorff distance in group-wise matching. In general, Euclidean distance is usually used to measure the similarity between samples, based on which a unidirectional network can be constructed (represented by using a connectivity matrix **P**). Since the Euclidean distance can only reflect the magnitude of similarity between two samples, it cannot represent the underlying dynamic structure information in data, which could lead to sub-optimal learning performance in subsequent learning tasks. Recently, a probabilistically motivated effective distance is proposed, which is shown reliable in predicting disease arrival times in the context of global air-traffic-mediated epidemics [28], by considering the dynamic structure information of data.

Regardless of the structural complexity of the underlying network, the key idea of effective distance is that the dynamic structure of data is dominated by a set of most probable paths, which can derived from the connectivity matrix P. Such hypothesis is analogous to the fact that the smallest resistor dominates a strongly heterogeneous electrical network with parallel conducting lines. Let P_{mn} ($0 \le P_{mn} \le 1$) denote the fraction of travelers that leave node n and arrive at node m. Then, the effective distance E_{mn} from a node n to a connected node m is defined as

$$E_{mn} = (1 - \log P_{mn}) \ge 1 \tag{5}$$

The effective distance defined in Eq. (5) reflects the idea that a small fraction of traffic $n \rightarrow m$ is effectively equivalent to a large distance, and vice versa. It is worth noting that Eq. (5) defines a quasi-distance, which is generally asymmetric (i.e., $E_{mn} \neq E_{nm}$). Such asymmetric characteristic is similar to a road network of one-way streets, where the shortest distance from A to B may differ from the one from B to A. Hence, such asymmetry captures the effect that a randomly seeded disease in a peripheral node of the network has a higher probability of being transmitted to a wellconnected hub than vice versa. As shown in [28], by using effective distance, patterns that have complex spatiotemporal structure in the conventional geographic perspective turn into regular, wavelike solutions reminiscent of simple reaction-diffusion systems. This permits the definition of effective epidemic wave fronts, propagation speeds, and the reliable estimation of epidemic arrival times, based on the knowledge of the underlying connectivity

network. In other words, effective distance can reveal hidden pattern geometry of data, compared to conventional geographic distance. Intuitively, using such effective distance in feature selection methods may help in finding the most discriminative features in data. To the best of our knowledge, no previous studies have used such effective distance for feature selection.

3. Proposed method

In this section, we first propose a sparse representation-based method to compute the effective distance, and then develop three novel unsupervised filter-type feature selection methods by using such effective distance.

3.1. Sparse representation-based effective distance

As mentioned in Section 2, the effective distance relies on a bidirectional network, which is represented by an asymmetric connectivity matrix P (i.e., the magnitude of P_{mn} is not necessarily equal to that of P_{nm}). To obtain such connectivity matrix, we resort to sparse representation that is proven robust to noisy, and help eliminate the negative effect of noisy features for subsequent feature selection process [50,51,53,54].

As an extension to traditional signal representations such as Wavelet and Fourier representations, sparse representation has been applied extensively in pattern recognition and signal processing [50,51,54–58]. Let $X = [x_1, \dots, x_N] \in \mathbb{R}^{d \times N}$ denote the data matrix, where $x_n \in \mathbb{R}^d$ represents a sample with d-dimensional features. Specifically, sparse representation aims to represent each x using as fewer entries of x as possible, which can be expressed formally as follows:

$$\min_{\mathbf{p}} ||\mathbf{p}||_{0}, \text{ s. t. } \mathbf{x} = \mathbf{X}\mathbf{p} \tag{6}$$

where $p \in \mathbb{R}^N$ is the coefficient vector, and $||p||_0$ is the pseudo- l_0 norm denoting the number of non-zero components in p. However, to find the sparsest solution of Eq. (6) is NP-hard, and it can be approximately solved by the following:

$$\min_{\mathbf{s}} \|\mathbf{p}\|_{1}, \text{ s. t. } \mathbf{x} = \mathbf{X}\mathbf{p} \tag{7}$$

where $\|\boldsymbol{p}\|_1$ is the l_1 norm of \boldsymbol{p} . It has been proven that the solution of l_1 norm minimization problem is equal to that of l_0 norm minimization problem, provided that the solution \boldsymbol{p} is sparse enough [59,60]. The problem in Eq. (7) can be solved by standard linear programming [56].

Recently, researchers in [61] construct a sparse reconstructive weight matrix based on a Modified Sparse Representation (MSR) framework, and show such a matrix helps to obtain more compact representation of data. Specifically, a sparse reconstructive weight vector \mathbf{p}_i for each \mathbf{x}_i can be obtained by solving the following modified l_1 minimization problem [56]:

$$\min_{\boldsymbol{p}_i} \|\boldsymbol{p}_i\|_1 \text{s. t. } \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{p}_i, 1 = 1^T \boldsymbol{p}_i$$
(8)

where $\mathbf{p}_i = [p_{i,1}, \dots, p_{i,i-1}, 0, p_{i,i+1}, \dots p_{i,N}]^T$ is an N-dimensional vector in which the i-th element is equal to zero implying that \mathbf{x}_i is removed from \mathbf{X} . The element $p_{i,j} (j \neq i)$ denotes the contribution of each \mathbf{x}_j to reconstruct \mathbf{x}_i , and $1 \in \mathbb{R}^N$ is a vector of all ones.

For each sample \mathbf{x}_i , we can compute the reconstructive weight vector \mathbf{p}_i , and then get the sparse reconstructive weight matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$:

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_N]^T \tag{9}$$

where $p_i(i=1,\cdots,N)$ is the optimal solution of Eq. (8). It is worth noting that the discriminative information can be naturally

preserved in the matrix P, even if no class label information is used. The reason is that the non-zero entries in p_i usually correspond to the samples from the same class, which implies that p_i may help to distinguish that class from the others. After obtaining the reconstruction weight matrix P through Eq. (9), the l_1 graph including both graph adjacency structure and affinity weights matrix can be simultaneously determined from P.

In many real-world problems, the constraint $x_i=Xp_i$ does not always hold. To overcome this problem, two modified objective functions are presented [50,61]. The first one is as follows:

$$\min_{\boldsymbol{p}_i} \|\boldsymbol{p}_i\|_1, \text{ s. t. } \|\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{p}_i\| < \delta, 1 = 1^T \boldsymbol{p}_i$$
(10)

where δ is the error tolerance. It can be seen that the optimal solution of Eq. (10) reflects some intrinsic geometric properties, e.g. invariant to translation and rotation. The second extension can be expressed as follows:

$$\min_{\begin{bmatrix} \boldsymbol{p}_i^T & \boldsymbol{t}_i^T \end{bmatrix}^T} \left\| \begin{bmatrix} \boldsymbol{p}_i^T & \boldsymbol{t}_i^T \end{bmatrix}^T \right\|_1, \text{ s. t.} \begin{bmatrix} \boldsymbol{x}_i \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{I} \\ 1^T & 0^T \end{bmatrix} \begin{bmatrix} \boldsymbol{p}_i \\ \boldsymbol{t}_i \end{bmatrix}$$
(11)

where t_i is a d-dimensional vector incorporated as a reconstructive compensation term. The optimal solution of Eq. (11) is also invariant to translations, but the invariance to rotation and re-scaling does not rigorously hold. In this way, we can obtain a bidirectional connectivity matrix P via Eq. (9).

By using the connectivity matrix P constructed via sparse representation, we can compute the effective distance matrix $ED = \left\{ED_{ij}\right\}_{i,j=1}^{N}$ according to Eq. (5). With this new distance, we can obtain an effective distance-based similarity matrix $ES = \left\{ES_{ij}\right\}_{i,j=1}^{N}$ that defines the similarity among samples. Here, the element ES_{ij} is defined as follows

$$ES_{ij} = e^{-\frac{ED_{ij}^2}{\lambda}} \tag{12}$$

where λ is a bandwidth parameter. Our approach is based on the idea of replacing conventional Euclidean distance by a measure of effective distance derived from the underlying bi-directional connectivity network. On the other hand, the effective distance-based similarity matrix **ES** actually defines a graph structure for data, where one node denotes a specific sample and a weighted edge is represented by an entry in matrix **ES**.

3.2. Effective distance-based feature selection

In the following, we introduce our effective distance-based feature selection methods, including 1) Effective Distance-based Laplacian Score (EDLS), 2) Effective Distance-based Sparsity Score-1 (EDSS-1), and 3) Effective Distance-based Sparsity Score-2 (EDSS-2).

Given an effective distance-based similarity matrix ES with element ES_{ij} defining the similarity between the i-th and the j-th samples, the effective distance-based Laplacian Score of the r-th feature, denoted as $EDLS_r$ that should be minimized, is computed as follows:

$$EDLS_{r} = \frac{\sum_{i=1}^{N} (f_{ri} - f_{rj})^{2} ES_{ij}}{\sum_{i=1}^{N} (f_{ri} - \mu_{r})^{2} D_{ii}}$$
(13)

where **D** is a diagonal matrix and $D_{ii} = \sum_{j=1}^{N} ES_{ij}$, and μ_r is the mean of the r-th feature among all samples.

Similarly, the effective distance-based Sparsity Score-1 (EDSS-1) and the effective distance-based Sparsity Score-2 (EDSS-2) of the r-th feature (denoted as $EDSS_r^1$ and $EDSS_r^2$, respectively) are defined as follows:

$$EDSS_r^1 = \sum_{i=1}^N \left(f_{ri} - \sum_{j=1}^N ES_{ij} f_{rj} \right)^2 = \mathbf{f}_r^T (\mathbf{I} - \mathbf{ES} - \mathbf{ES}^T + \mathbf{ESES}^T) \mathbf{f}_r$$
(14)

$$EDSS_{r}^{2} = \frac{\sum_{i=1}^{N} \left(f_{ri} - \sum_{j=1}^{N} ES_{ij} f_{rj} \right)^{2}}{\frac{1}{N} \sum_{i=1}^{N} \left(f_{ri} - \mu_{r} \right)^{2}} = \frac{\mathbf{f}_{r}^{T} \left(\mathbf{I} - \mathbf{ES} - \mathbf{ES}^{T} + \mathbf{ESES}^{T} \right) \mathbf{f}_{r}}{\mathbf{f}_{r}^{T} \left(\mathbf{I} - \frac{1}{N} \mathbf{1} \mathbf{1}^{T} \right) \mathbf{f}_{r}}$$
(15)

where ES_{ij} is the entry of the effective distance-based similarity matrix ES. Similar to Sparsity Score [49], by minimizing $EDSS_r^1$ ($EDSS_r^2$), features that can best respect the pre-defined effective distance-based graph structure are regarded as being more discriminative. In Algorithm 1, we list the details of our proposed effective distance-based feature selection algorithm.

Algorithm 1. Learning Algorithm for Effective Distance-based Feature Selection.

Input: The training data $X = \{x_i\}_{i=1}^N$. **Initialize**: The bandwidth parameter λ .

Step 1. Construct the sparse reconstruction coefficient matrix $P \in \mathbb{R}^{N \times N}$, and normalize each column of P to [0, 1].

Step 2. Compute the effective distance matrix $ED = 1 - \log(P)$.

Step 3. Construct the effective distance-based similarity matrix **ES**, with element ES_{ii} is computed according to Eq. (12).

Step 4. Using **S**, we compute the effective distance-based Laplacian Score (EDLS), effective distance-based Sparsity Scores (i.e., EDSS-1, and EDSS-2) via Eqs. (13)—(15).

Step 5. Rank features according to EDLS, EDSS-1 and EDSS-2 in ascending order, respectively.

Output: Feature ranking lists.

3.3. Computational complexity analysis

Now we analyze the computational complexity of Algorithm 1. There are three main steps in Algorithm 1: 1) Step 1 constructs the bi-directional connectivity matrix P via sparse representation using Eq. (10) or Eq. (11), requiring $O(N^2)$ operations given N data points; 2) Step computes the effective distance matrix ED using Eq. (5), which needs $O(N^2)$ operations; 3) Step 3 calculates the effective distance-based similarity matrix ES, which needs $O(N^2)$ operations; 4) Step 4 evaluates d features requiring O(dN) operations for EDLS, and $O(dN^2)$ operations for both EDSS-1 and EDSS-2; (5) Step 5 ranks d features that needs $O(d \log d)$ operations. Hence, the overall time complexity of Algorithm 1 is $O(d \max(N, \log d))$ for the proposed EDLS method, and $O(d \max(N^2, \log d))$ for both EDSS-1 and EDSS-2 methods.

4. Experiments

4.1. Data sets

First, we evaluate our proposed methods (i.e., EDLS, EDSS-1, and EDSS-2) on ten data sets from UCI machine learning repository [62]. The proposed method is also used for recognition and tested on two well-known face image data sets, i.e. the Yale Face [63] and the ORL Face [64]. The statistics of data sets used in our experiments are summarized in Table 1.

Specifically, the Yale face data set contains 165 gray-scale images of 15 persons, 11 images per person. Images for each person are taken at different facial expressions or configurations. The size of each cropped image is 32×32 pixels, with 256 grey levels per pixel. Thus, each image is represented by a feature vector by a

 Table 1

 Characteristics of data sets used in the experiments.

Data set	# class	# instance	# feature
Anneal	5	898	90
Breast Tissue	6	106	9
Colic	2	368	60
Ecoli	8	336	7
Glass	6	214	9
Hepatitis	2	155	19
Sonar	2	208	60
Wdbc	2	569	30
Wine	3	178	13
Zoo	7	101	17
Yale Face	15	165	1024
ORL Face	40	400	1024

1024-dimensionality. No further preprocessing is done. Fig. 2 (a) shows the 11 images for one of these 15 persons. The clustering experiments are based on the K-means algorithms. This process along with the corresponding classifier learning is repeated 10 times. The ORL face data set contains 400 gray-scale images of 40 distinct persons, 10 different images per person. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement), with sample images shown in Fig. 2(b). Similar to the Yale face database, we resize each image in the ORL face data set to 32×32 pixels.

4.2. Experiment design

We empirically evaluate the performance of our proposed methods (i.e., EDLS, EDSS-1, and EDSS-2) in both unsupervised and supervised learning contexts. Specifically, there are two groups of experiments, where the first one is to use feature selection for unsupervised clustering tasks, and the second one is to perform supervised classification tasks by using different feature selection methods. It is worth noting that the bandwidth parameter λ in Eq. (2) for Laplacian Score (LS), as well as that in Eq. (12) our proposed EDLS method, is set as the mean distance among all training samples without any extra explanation. We will further investigate the influence of this parameter on our method in Section 4.5.

4.2.1. Clustering experiment setting

For the clustering experiments, we use the K-means algorithms to perform clustering. To be specific, for each given number K, K classes were randomly selected from the face data set. This process was repeated 10 times (except for K is equal to the real cluster number) and the average performance was computed. For each test (given K classes), specific unsupervised feature selection

algorithms (e.g., EDLS, EDSS-1 and EDSS-2) are used to select the features. The K-means was then performed in the selected feature subspace. Again, the K-means is repeated 10 times with different initializations and the best result in terms of the objective function of K-means was recorded.

By utilizing different feature selection algorithms, we select a feature subset that is most useful for discriminating different classes. Then clustering is performed in such a subspace. To be specific, firstly, a specific feature selection algorithm (e.g., Var, LS, SS-1, SS-2) are used to select the most discriminative features. Secondly, we utilize the K-means algorithm to perform clustering with the selected feature subspace. The K-means clustering process is repeated for 10 times with different initializations and the best results in terms the objective function of K-means was recorded. It is worth noting the initialization is the same for different algorithms for fair comparison. Meanwhile, we have compared the proposed methods with Baselines (i.e., the clustering results without any feature selection procedure).

By comparing the obtained label of each data points of K-means algorithm with that provided by the data corpus, the clustering result can be evaluated. The F-Score metric [65] is used to measure the clustering performance. Given a clustering result, the F-Score is defined as follows:

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
(16)

where Precision and Recall are two measure criteria. Here, the Precision and the Recall are defined as follows:

$$Precision = \frac{N_1}{N_1 + N_2} \tag{17}$$

$$Recall = \frac{N_1}{N_1 + N_3} \tag{18}$$

where N_1 is the number of sample pairs which are clustered correctly, N_2 is the number of sample pairs which belong to the different classes but are clustered into the same class, and N_3 is the number of sample pairs which belong to the same class but are clustered into different classes.

4.2.2. Classification experiment setting

In the classification experiments, we adopt a 5-fold cross validation strategy to compute the classification accuracy. To be specific, we first partition the whole data set into 5 subsets (each subset with roughly equal size), and each time one of these subsets is utilized as the test set while the other 4 subsets are combined together to be the training set. In addition, to determine the number of optimal feature dimensions in each fold, we further perform feature selection via an inner cross-validation on training data (i.e., another 5-fold cross-validation is performed on training data). Then, the mean and the standard deviation of classification



(a) Yale Face



(b) ORL Face

Fig. 2. Sample images in (a) Yale Face data set, and (b) ORL Face data set (a) Yale Face (b) ORL Face.

accuracies on the test set using such optimal feature subset are reported. Meanwhile, the classification accuracy on the test set without any feature selection (i.e. using original features) is used as the Baseline. Finally, the 1-Nearest-Neighborhood (1-N-N) classifier is utilized to perform classification by using those selected features.

4.3. Results of clustering experiments

In this group of experiments, we validate the efficacy of proposed methods (i.e., EDLS, EDSS-1, EDSS-2) in an unsupervised problem setting, in comparison to several well-known unsupervised feature selection methods, including Variance (VS) [8]. Laplacian Score (LS) [48], Sparsity Score-1 (SS-1) and Sparsity Score-2 (SS-2) [49]. The performance of a specific feature selection method is measured by the F-score based on the selected features on the training data. Specifically, for a specific feature selection method, we first select the first *m* features from the ranking list of features generated by corresponding algorithms, where m is the desired number of selected features specified as $m = \{1, 2, ..., d\}$ in the experiments. Then, we report the highest classification result as well the number of selected features. It is worth noting that the cluster number in this set of experiments is given in accordance with the true class number of a specific data set, if without extra explanations.

In Table 2, we report the clustering results using K-means algorithm, where the Baseline is achieved by using all features (i.e., without any feature selection process). The meaning of the symbols in the term 'a \pm b' is as follows: 'a' and 'b' denote the mean and the variance of F-score among 10 times, respectively. Note that the best results are shown in boldface. Then, in Fig. 3, we plot the clustering result vs. the number of selected features on the Breast Tissue and the Wine data sets, achieved by our proposed effective distance-based methods (i.e., EDLS, EDSS-1, and EDSS-2) and their conventional counterparts (i.e., LS, SS-1, and SS-2).

From Table 2, we can see four main points. First, our proposed effective distance-based feature selection methods (i.e., EDLS, EDSS-1, and EDSS-2) usually achieve the overall better performances than their conventional counterparts (i.e., LS, SS-1, and SS-2). For example, on the Zoo data set, the F-score achieved by the proposed EDSS-2 approach is 0.7784, while the best accuracy of SS-2 is only 0.7334. Second, our proposed EDSS-2 method often performs better than EDLS and EDSS-1. For example, EDSS-2 achieves the best results in 7 out of 10 data sets. It indicates that the effective distance is important in guiding the process of feature selection. From Fig. 3, we can see again that our proposed effective distance-based methods (i.e., EDLS, EDSS-1, and EDSS-2) achieve the overall best performances using smaller number of selected features, in comparison to LS, SS-1 and SS-2. For example, on the Breast Tissue data set, our proposed EDLS method achieves much higher F-score than LS.

 Table 2

 Clustering results of different unsupervised feature selection methods.

In addition, we report the clustering results of image recognition on both the Yale and the ORL face data sets using K-means algorithm in Fig. 4. From Fig. 4, one can observe that our proposed methods (i.e., EDLS, EDSS-1, and EDSS-2) usually outperform the compared methods in two image clustering tasks. Especially, as shown in Fig. 4(a), EDSS-2 achieves much higher F-score than Baseline, VS, LS, SS-1 and SS-2. At the same time, the proposed EDLS, EDSS-1 and EDSS-2 methods consistently outperform their conventional counterparts (i.e., LS, SS-1, and SS-2), respectively.

Furthermore, we investigate the influence of cluster numbers on the clustering performance. In Fig. 5, we report the results on the ORL Face data set with different cluster numbers. As can be seen from Fig. 5, the proposed EDLS, EDSS-1 and EDSS-2 methods consistently outperform their conventional counterparts (i.e., LS, SS-1, and SS-2) with the increase of the cluster number. The reason may be that feature selection methods based on effective distance can help find more discriminative features than methods using conventional distance measures.

4.4. Results of classification experiments

In the second group of experiments, we evaluate our proposed methods in classification tasks. For fair comparison, we first compare our methods with several unsupervised feature selection methods (including VS, LS, SS-1, and SS-2), with results reported in Table 3. In Fig. 6, we further plot the curves of classification accuracy vs. number of selected features on two UCI data sets.

From Table 3, we can see that the performances of EDLS, EDSS-1 and EDSS-2 are superior to those of Var, LS, SS-1, SS-2 and Baseline in most data sets. It is worth noting that the proposed EDSS-2 method achieves much higher averaged accuracies than those of the other seven methods. From Fig. 6, one can see that the proposed EDLS, EDSS-1 and EDSS-2 usually outperform LS, SS-1, and SS-2, respectively. It validates the efficacy of the proposed methods using effective distance measure.

5. Discussion

In this work, we propose a sparse representation-based effective distance computation method. The main advantage of our method is that such distance measure can reflect the global and dynamic property of data. Also, we further develop three effective distance-based unsupervised feature selection methods, by replacing traditional Euclidean distance with the proposed effective distance. It is worth noting that, as a general similarity measure metric, the proposed effective distance can also be used in many other domains, such as dimension reduction [66,67] and graph-based learning methods [27]. In what follows, we will first compare our methods with several state-of-the-art methods, and then discuss the influence of parameter on the learning performance.

Data set	Baseline	VS	LS	SS-1	SS-2	EDLS	EDSS-1	EDSS-2
Anneal	0.5549 ± 0.1269	0.6226 ± 0.1269	0.7481 ± 0.1269	0.7496 ± 0.1269	0.7563 ± 0.1621	0.7498 ± 0.1269	0.7563 ± 0.1591	0.7627 ± 0.1621
Breast Tissue	0.3562 ± 0.0704	0.3583 ± 0.0067	0.4364 ± 0.0458	0.3577 ± 0.0564	0.3562 ± 0.0034	0.4542 ± 0.0458	0.4477 ± 0.0067	$\bf 0.4602 \pm 0.0033$
Colic	0.5397 ± 0.0636	0.5633 ± 0.0541	0.5615 ± 0.0512	0.6725 ± 0.0636	0.6568 ± 0.0526	0.6823 ± 0.0512	0.6930 ± 0.0456	$\bf 0.7018 \pm 0.0510$
Ecoli	0.5199 ± 0.0347	0.5199 ± 0.0648	0.5253 ± 0.0529	0.5199 ± 0.0648	0.5207 ± 0.0529	$\bf 0.5352 \pm 0.0529$	0.5346 ± 0.0529	0.5299 ± 0.0648
Glass	0.4324 ± 0.0497	0.4690 ± 0.0422	0.4613 ± 0.0365	0.4584 ± 0.0365	0.4525 ± 0.0365	$\bf 0.4868 \pm 0.0422$	0.4575 ± 0.0387	0.4808 ± 0.0462
Hepatitis	0.6858 ± 0.0663	0.7173 ± 0.0663	0.7100 ± 0.0663	0.6982 ± 0.0499	0.8002 ± 0.0717	0.7045 ± 0.0542	0.8013 ± 0.0717	$\textbf{0.8015} \pm \textbf{0.0500}$
Sonar	0.5021 ± 0.0213	0.5570 ± 0.0213	0.5208 ± 0.0150	0.5237 ± 0.0364	0.6070 ± 0.0364	0.5998 ± 0.0175	0.6031 ± 0.0213	0.6092 ± 0.0169
Wdbc	0.7878 ± 0.0822	0.7878 ± 0.0859	0.7878 ± 0.0971	0.7878 ± 0.0859	0.8412 ± 0.0217	0.8269 ± 0.0717	0.7878 ± 0.0084	0.8287 ± 0.0098
Wine	0.5861 ± 0.0843	0.5941 ± 0.0843	0.5879 ± 0.0243	0.5871 ± 0.0064	0.5942 ± 0.0243	0.6442 ± 0.0243	0.6647 ± 0.0843	0.6862 ± 0.0064
Zoo	0.6856 ± 0.1012	0.6856 ± 0.0927	0.7209 ± 0.0862	0.7290 ± 0.0798	0.7334 ± 0.1165	0.7769 ± 0.0862	0.7556 ± 0.0927	$\bf 0.7784 \pm 0.0890$
Average	0.5651 ± 0.0701	0.5875 ± 0.0645	0.6060 ± 0.0602	0.6083 ± 0.0607	0.6319 ± 0.0578	0.6461 ± 0.0573	0.6516 ± 0.0637	$\bf 0.6639 \pm 0.0499$

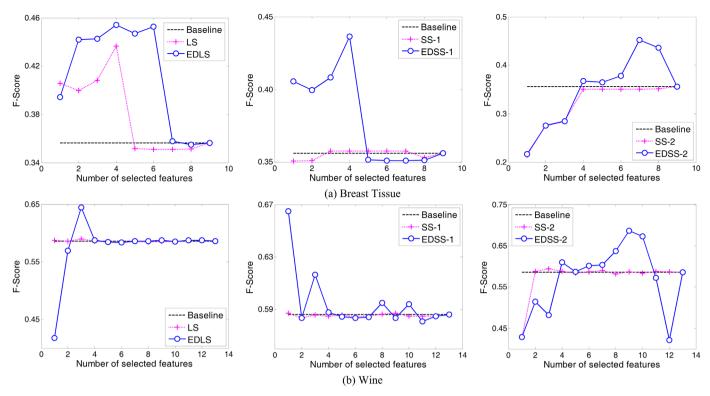


Fig. 3. Clustering results achieved by effective distance-based methods and their conventional counterparts on (a) Breast Tissue data set, and (b) Wine data set.

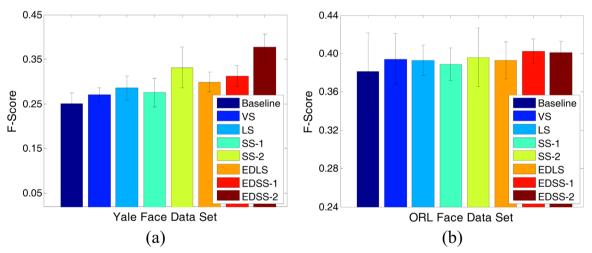


Fig. 4. Clustering results achieved by different methods on (a) Yale Face data set (15 clusters), and (b) ORL Face data set (40 clusters).

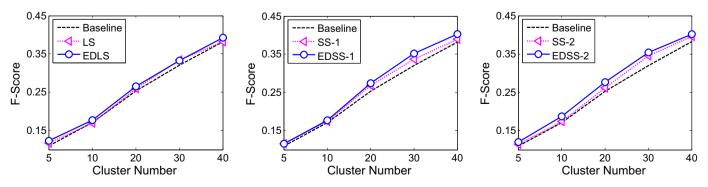


Fig. 5. Clustering results using different cluster numbers on ORL Face data set.

Table 3Classification results of different unsupervised feature selection methods.

Data set	Baseline	VS	LS	SS-1	SS-2	EDLS	EDSS-1	EDSS-2
Anneal	0.8875 ± 0.0025	0.8960 ± 0.0101	0.8973 ± 0.0066	0.8955 ± 0.0095	0.9540 ± 0.0207	0.9228 ± 0.0347	0.9549 ± 0.0446	0.9848 ± 0.0112
Breast Tissue	0.5308 ± 0.0532	0.5308 ± 0.0496	0.5308 ± 0.0462	0.5346 ± 0.0496	0.5308 ± 0.0465	0.6385 ± 0.0411	0.6192 ± 0.0445	$\bf 0.6423 \pm 0.0396$
Colic	0.6424 ± 0.0364	0.6478 ± 0.0216	0.6674 ± 0.0296	0.7283 ± 0.0452	0.6511 ± 0.0334	$\textbf{0.7543} \pm \textbf{0.0351}$	0.7391 ± 0.0548	0.7380 ± 0.0568
Ecoli	0.7265 ± 0.0155	0.7265 ± 0.0299	0.7265 ± 0.0299	0.7265 ± 0.0299	0.7265 ± 0.0284	0.7398 ± 0.0227	0.7398 ± 0.0227	$\textbf{0.7482} \pm \textbf{0.0145}$
Glass	0.6990 ± 0.0156	0.6990 ± 0.0230	0.6990 ± 0.0230	0.6990 ± 0.0230	0.6990 ± 0.0311	$\textbf{0.7181} \pm \textbf{0.0155}$	$\textbf{0.7181} \pm \textbf{0.0155}$	0.7181 ± 0.0155
Hepatitis	0.6727 ± 0.0351	0.6727 ± 0.0571	0.6883 ± 0.0677	0.6727 ± 0.0571	$\bf 0.7922 \pm 0.0289$	0.7636 ± 0.0611	0.7662 ± 0.0539	0.7948 ± 0.0142
Sonar	0.7359 ± 0.0108	0.7398 ± 0.0129	0.7398 ± 0.0143	0.7417 ± 0.0169	0.7534 ± 0.0200	0.7748 ± 0.0143	0.7612 ± 0.0279	$\textbf{0.7806} \pm \textbf{0.0323}$
Wdbc	0.9070 ± 0.0116	0.9070 ± 0.0138	0.9077 ± 0.0136	0.9070 ± 0.0138	0.9190 ± 0.0116	$\bf 0.9268 \pm 0.0123$	0.9077 ± 0.0136	0.9155 ± 0.0050
Wine	0.7023 ± 0.0532	0.7045 ± 0.0431	0.7045 ± 0.0498	0.7045 ± 0.0498	0.7409 ± 0.0208	0.8045 ± 0.0858	0.8500 ± 0.0913	$\bf 0.8659 \pm 0.0520$
Zoo	0.9306 ± 0.0175	0.9347 ± 0.0271	0.9388 ± 0.0289	0.9388 ± 0.0289	0.9429 ± 0.0327	$\bf 0.9429 \pm 0.0238$	0.9388 ± 0.0316	0.9306 ± 0.0208
Average	0.7435 ± 0.0251	0.7459 ± 0.0288	0.7500 ± 0.0310	0.7549 ± 0.0323	0.7710 ± 0.0274	0.7986 ± 0.0346	0.7995 ± 0.0400	$\textbf{0.8119} \pm \textbf{0.0262}$

5.1. Comparison with state-of-the-art methods

We compare our methods with several state-of-the-art supervised feature selection methods: 1) Fisher Score (FS) [8], 2) Constraint Score (CS) [17], 3) SVM-RFE (RFE) [68] Fisher-Markov selector with linear kernel (LFS) [35]. The experimental results are shown in Fig. 7. As can be seen from Fig. 7, our proposed effective distance-based feature selection methods perform better than the compared methods in 6 out of 10 data sets. These results demonstrate that our proposed unsupervised feature selection methods can achieve comparative or better results, compared with several supervised methods.

5.2. Influence of parameter

In our proposed effective distance computation method, there is a bandwidth parameter to be tuned (i.e., λ) in Eq. (12). In this section, we discuss the influences of parameters on the clustering performances of our methods. In Fig. 8, we plot the curves of classification accuracies as a function of the bandwidth parameter (i.e., λ) on both the Breast Tissue and the Wine data sets.

From Fig. 8(a)-(b), we can clearly see the performances of the proposed methods (i.e., EDLS, EDSS-1, and EDSS-2) are generally stable with respect to the different values of λ . At the same time, the proposed EDSS-2 consistently outperforms EDLS and EDSS-1 with the increase of parameter value. These results illustrate that our proposed effective distance-based methods are not sensitive to the bandwidth parameter.

6. Conclusion

In this paper, we propose a set of effective distance-based feature selection methods, where a probabilistically motivated effective distance is used to measure the similarity between a pair of samples. To be specific, we first propose a sparse representation-based algorithm to compute the effective distance. Then, three new filter-type unsupervised feature selection methods using effective distance are developed, including an effective distance-based Laplacian Score (EDLS), and two effective distance-based Sparsity Scores (EDSS-1, and EDSS-2). Experimental results

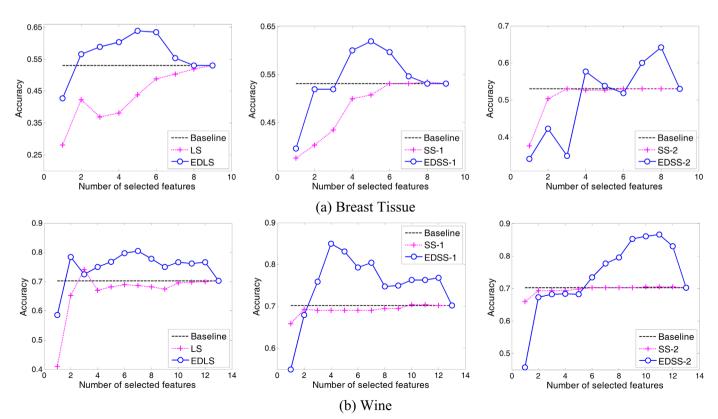


Fig. 6. Classification results using conventional distance and effective distance on (a) Breast Tissue data set, and (b) Wine data set.

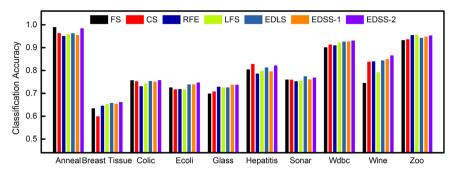


Fig. 7. Comparision with supervised feature selection methods on ten UCI data sets.

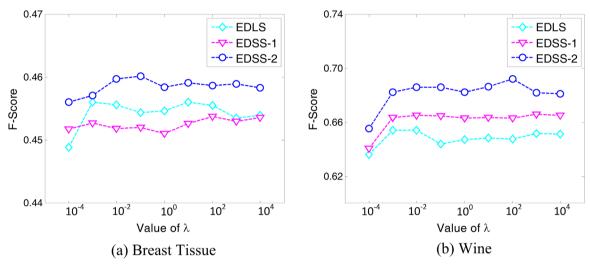


Fig. 8. Clustering results by varying the bandwidth parameter on (a) Breast Tissue, and (b) Wine data sets.

of clustering and classification tasks on a series of benchmark data sets demonstrate the efficacy of our proposed methods, compared with methods using conventional Euclidean distance. It is worth noting that the proposed sparse representation-based effective distance computation method can also be used in dimension reduction and spectral clustering, which will be our future work.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Nos. 61422204, 61473149, 61402320), the Jiangsu Natural Science Foundation for Distinguished Young Scholar (No. BK20130034), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20123218110009), the NUAA Fundamental Research Funds (No. NE2013105), and the Natural Science Foundation of Shandong Province of China (No. ZR2013F0029).

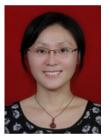
References

- R. He, W.S. Zheng, T.N. Tan, Z.A. Sun, Half-quadratic-based iterative minimization for robust sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 261–275.
- [2] D. Dernoncourt, B. Hanczar, J.D. Zucker, Analysis of feature selection stability on high dimension and small sample data, Comput. Stat. Data Anal. 71 (2014) 681–693.
- [3] G. Wang, J. Ma, S.L. Yang, An improved boosting based on feature selection for corporate bankruptcy prediction, Expert. Syst. Appl. 41 (2014) 2353–2361.
- [4] M. Liu, D. Zhang, D. Shen, View-centralized multi-atlas classification for Alzheimer's disease diagnosis, Hum. Brain Mapp. 36 (2015) 1847–1865.
- [5] M. Liu, D. Zhang, Pairwise constraint-guided sparse learning for feature selection, IEEE Trans. Cybern. 46 (2016) 298–310.
- [6] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, IEEE Trans. Image Process. 22 (2013) 363–376.

- [7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 33 (2011) 856–867.
- [8] A.R. Webb, Statistical Pattern Recognition, John Wiley & Sons, 2003.
- 9] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4–37.
- [10] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
- [11] J. Jaeger, R. Sengupta, W.L. Ruzzo, Improved gene selection for classification of microarrays, in: Proceedings of Pacific Symposium on Biocomputing, Lihue, Hawaii, 2003, pp. 53–64.
- [12] E.R. Hancock, Z. Zhang, Kernel entropy-based unsupervised spectral feature selection, Int. J. Pattern Recognit. Artif. Intell. 26 (2012) 1260002.
- [13] P. Singh, V. Laxmi, M.S. Gaur, Near-optimal geometric feature selection for visual speech recognition, Int. J. Pattern Recognit. Artif. Intell. 27 (2013) 1350026.
- [14] S. Gao, C. Zu, D. Zhang, Learning mid-perpendicular hyperplane similarity from cannot-link constraints, Neurocomputing 113 (2013) 195–203.
- [15] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 40–51.
- [16] Z. Lu, Y. Peng, Exhaustive and efficient constraint propagation: a graph-based learning approach and its applications, Int. J. Comput. Vis. 103 (2013) 306–325.
- [17] D.Q. Zhang, S.C. Chen, Z.H. Zhou, Constraint score: a new filter method for feature selection with pairwise constraints, Pattern Recognit. 41 (2008) 1440–1451.
- [18] D. Sun, D.Q. Zhang, Bagging constraint score for feature selection with pairwise constraints, Pattern Recognit. 43 (2010) 2106–2118.
- [19] B. Hariharan, S.V.N. Vishwanathan, M. Varma, Efficient max-margin multi-label classification with applications to zero-shot learning, Mach. Learn. 88 (2012) 127–155.
- [20] C. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zeroshot learning of object categories, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 453–465.
- [21] M. Liu, D. Zhang, S. Chen, H. Xue, Joint binary classifier learning for ECOC-based multi-class classification, IEEE Trans. Pattern Anal. Mach. Intell. (2015) 1.
- [22] M. Liu, D. Zhang, S. Chen, Attribute relation learning for zero-shot classification, Neurocomputing 139 (2014) 34–46.
- [23] Q. Liu, Y. Yang, R. Ji, Y. Gao, L. Yu, Cross-view down/up-sampling method for multiview depth video coding, IEEE Signal Process. Lett. 19 (2012) 295–298.
- [24] H. Elghazel, A. Aussem, Unsupervised feature selection with ensemble learning, Mach. Learn. 98 (2015) 157–180.

- [25] A.K. Farahat, A. Ghodsi, M.S. Kamel, Efficient greedy feature selection for unsupervised learning, Knowl. Inf. Syst. 35 (2013) 285–310.
- [26] J. Zhang, D. Zhang, A novel ensemble construction method for multi-view data using random cross-view correlation between within-class examples, Pattern Recognit. 44 (2011) 1162–1171.
- [27] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3D object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (2012) 4290–4303.
- [28] D. Brockmann, D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science 342 (2013) 1337–1342.
- [29] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Feature Extraction, Foundations and Applications, Springer, 2006.
- [30] H. Yoon, K. Yang, C. Shahabi, Feature subset selection and feature ranking for mutivariate time series, IEEE Trans. Knowl. Data Eng. 17 (2005) 1186–1198.
- [31] S. Nakariyakul, D.P. Casasent, An improvement on floating search algorithms for feature subset selection, Pattern Recognit. 42 (2009) 1932–1940.
- [32] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
- [33] D. Rodrigues, L.A.M. Pereira, R.Y.M. Nakamura, K.A.P. Costa, X.S. Yang, A. N. Souza, J.P. Papa, A wrapper approach for feature selection and optimumpath forest based on bat algorithm, Expert. Syst. Appl. 41 (2014) 2250–2258.
- [34] A.L. Blum, P. Langley, Selection of relevant features in machine learning, Artif. Intell. 97 (1994) 245–271.
- [35] Q.A. Cheng, H.B. Zhou, J. Cheng, The fisher-markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1217–1233.
- [36] N. Kwak, C.H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1667–1671.
- [37] L. Yu and H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proceedings of the 20th International Conference on Machine Learning, Washington DC, 2003, pp. 601-608.
- [38] K. Kira and L. Rendell, The feature selection problem: Traditional methods and a new algorithm, in: Proceedings of the 10th National Conference on Artificial Intelligence, 1992, pp. 129-134.
- [39] M. Liu, L. Miao, D. Zhang, Two-stage cost-sensitive learning for software defect prediction, IEEE Trans. Reliab. 63 (2014) 676–686.
- [40] C.X. Shang, M. Li, S.Z. Feng, Q.S. Jiang, J.P. Fan, Feature selection via maximizing global information gain for text classification, Knowl. Based Syst. 54 (2013) 298–309.
- [41] Y. Yang, X. Wang, Q. Liu, M. Xu, L. Yu, A bundled-optimization model of multiview dense depth map synthesis for dynamic scene reconstruction, Inf. Sci. 320 (2014) 306–319.
- [42] D. Ziou, T. Hamri, S. Boutemedjet, A hybrid probabilistic framework for content-based image retrieval with feature weighting, Pattern Recognit. 42 (2009) 1511–1519
- [43] A. Anand, G. Pugalenthi, P. Suganthan, Predicting protein structural class by SVM with class-wise optimized features and decision probabilities, Pattern Recognit. 253 (2008) 375–380.
- [44] Y. Li, B. Lu, Feature selection based on loss-margin of nearest neighbor classification. Pattern Recognit. 42 (2009) 1914–1921.
- [45] H. Liu, J. Sun, L. Liu, H. Zhang, Feature selection with dynamic mutual information, Pattern Recognit. 42 (2009) 1330–1339.
- [46] Y. Yang, X. Wang, T. Guan, J. Shen, L. Yu, A multi-dimensional image quality prediction model for user-generated images in social networks, Inf. Sci. 281 (2014) 601–610.
- [47] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
- [48] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. (2005).
- [49] M. Liu, D. Zhang, Sparsity score: a novel graph-preserving feature selection method, Int. Journal. Pattern Recognit. Artif. Intell. 28 (2014) 1450009.
- [50] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.
- [51] K. Huang, S. Aviyente, Sparse representation for signal classification, Adv. Neural Inf. Process. Syst. (2006) 609–616
- Neural Inf. Process. Syst. (2006) 609–616. [52] Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai, 3-D object retrieval with Hausdorff distance learning, IEEE Trans. Ind. Electron. 61 (2014) 2088–2098.
- [53] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 15 (2006) 3736–3745.

- [54] T.T. Cai, A.R. Zhang, Sparse representation of a polytope and recovery of sparse signals and low-rank matrices, IEEE Trans. Inf. Theory 60 (2014) 122–132.
- [55] Y. Xie, W.S. Zhang, Y.Y. Qu, Y.H. Zhang, Discriminative subspace learning with sparse representation view-based model for robust visual tracking, Pattern Recognit. 47 (2014) 1383–1394.
- [56] S.S.B. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev. v43 (2001) 129–159.
- [57] J. Murray, K. Kreutz-Delgado, Visual recognition and inference using dynamic overcomplete sparse learning, Neural Comput. 19 (2007) 2301–2352.
- [58] G.Q. Shao, Y.P. Wu, A. Yong, X. Liu, T.D. Guo, Fingerprint compression based on sparse representation, IEEE Trans. Image Process. 23 (2014) 489–501.
- [59] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306.
- [60] R. Baraniuk, A lecture on compressive sensing, IEEE Signal Process. Mag. 24 (2007) 118–121.
- [61] L.S. Qiao, S.C. Chen, X.Y. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (2010) 331–341.
- [62] A. Frank, A. Asuncion, UCI Machine Learning Repository, ed: School of Information and Computer Science, University of California, 2007.
- [63] The Yale Face Database. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html).
- [64] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, Proc. Second. IEEE Workshop Appl. Comput. Vis. (1994) 138–142
- [65] K. Nigam, A. Mccallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM, Mach. Learn. 39 (2001) 103–134.
- [66] C. Boutsidis, A. Zouzias, M.W. Mahoney, P. Drineas, Randomized dimensionality reduction for k-means clustering, IEEE Trans. Inf. Theory 61 (2015) 1045.
- [67] S. Chen, D. Zhang, Semisupervised dimensionality reduction with pairwise constraints for hyperspectral image classification, IEEE Geosci. Remote. Sens. Lett. 8 (2011) 369–373.
- [68] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.



Mingxia Liu, received the B.S. degree, M.S. degree from Shandong Normal University, China, in 2003 and 2006, respectively, and Ph.D. degree from Nanjing University of Aeronautics and Astronautics in 2015, China. Her research interests include machine learning, pattern recognition, computer vision, and neuroimaging analysis.



Daoqiang Zhang, received the B.S. degree and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999 and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a Lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published over 100 scientific articles in refereed international journals such as IEEE Trans. Pattern Analysis and Machine Intelligence, Neuroimage, Pattern Recognition, Human Brain Mapping, IEEE Trans. Neural Networks; and conference proceedings such as

IJCAI, AAAI, MICCAI, ICDM. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).