Neurocomputing 259 (2017) 146–153

Contents lists available at ScienceDirect

Neurocomputing

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

Iterative sparsity score for feature selection and its extension for multimodal data

Chen Zu[a,1], Linling Zhu[b,1], Daoqiang Zhang[a,*]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[b] Patent Examination Cooperation Jiangsu Center of the Patent Office, Suzhou, China

ARTICLE INFO

Article history:
Received 12 February 2016
Revised 10 June 2016
Accepted 3 August 2016
Available online 21 February 2017

MSC:
00-01
99-00

Keywords:
Feature selection
Iterative sparse representation
Multi-modality
Alzheimer's disease
Clustering
Classification

ABSTRACT

As a key dimensionality reduction technique in pattern recognition, feature selection has been widely used in information retrieval, text classification and genetic data analysis. In recent years, structural information contained in samples for guiding feature selection has become a new hot spot in machine learning field. Although tremendous feature selection methods have been developed, less important features are still used to construct the structure in those conventional structure based feature selection approaches. In this paper, we propose a new filter-type feature selection method called iterative sparsity score, which is independent of any learning algorithm. The proposed method can preserve the structural information by sparse representation, which can be efficiently solved by a ℓ_1-norm minimization problem. To exclude data noise, at one time we discard last m features and iteratively optimize the ℓ_1-norm minimization problem. We perform clustering and classification experiments on numerous bench mark datasets. Furthermore, its extension for multimodal data is also developed. We adopt the multi-modality alzheimer's disease data for classification to evaluate the extended method. The experimental results show the effectiveness of our proposed methods compared with several popular feature selection approaches.

1. Introduction

With the rapid development of technology, big data have emerged with large dimensions or huge numbers. For example, high resolution face pictures include a lot of information with high dimensions, but at the same time the high dimensions limit the utilization in practical applications. One 3-D magnetic resonance image contains tremendous features, while the number of samples is very small in medical image analysis [1–3]. For biomedical research, thousands of genes are in dozens of samples of microarray databases [4]. Those applications are typical small sample size problems, which mean there is small number of samples with high dimensionality in those data. It is very difficult to use machine learning methods for learning on those data. Directly adopting the high dimensional data for classification or clustering is time-consuming [5,6]. The features extracted from raw data always contain noises, which affect its true distribution. And those noisy features will reduce the performance of learning.

Dimensionality reduction which can be divided into features selection and feature extraction is used for reducing the number of random variables under consideration [7,8] in machine learning and pattern recognition fields [9–12]. Feature selection methods [13–15] try to find a subset of features and feature extraction [16,17] approaches transform the data in the high-dimensional space to a new feature space with fewer dimensions. Filter [18] and wrapper [19] strategies are widely used in feature selection approaches. Filter based methods evaluate features via intrinsic properties (e.g., information gain) of the data. Wrapper techniques select features on which the learning process can achieve the predefined goal. In our paper, we currently focus on filter-type feature selection methods.

There are several popular filter-type feature selection methods, such as variance (Var) [20], Fisher Score (FS) [20], Laplacian Score (LS) [21], and Sparsity Score (SS) [22]. Among those filter-type feature selection approaches, Sparsity Score (SS) based on ℓ_1-norm minimization can preserve the structure of a predefined graph. However, SS involves all the features constructing the predefined graph. Obviously, the noisy features will embed irrelevant information into the graph. In this paper, we propose a novel iterative sparsity score method to solve the above problem. At a time, we discard last m features to construct the graph, and then iteratively

* Corresponding author.
E-mail addresses: chenzu@nuaa.edu.cn (C. Zu), zhu_linling@126.com (L. Zhu), dqzhang@nuaa.edu.cn (D. Zhang).
[1] Chen Zu and Linling Zhu are co-first authors

http://dx.doi.org/10.1016/j.neucom.2016.08.124
0925-2312/

solve the ℓ_1-norm minimization problem. The proposed method is compared with 4 popular filter-type feature selection approaches (including supervised and unsupervised). Besides, we also extend our proposed feature selection method to deal with multimodal problems. Experimental results on both clustering and multimodal classification demonstrate the effectiveness of our proposed algorithms.

The rest of the paper is organized as follows. The background is introduced and several typical filter-type feature selection methods are discussed in Section 2. In Section 3, we present the proposed iterative sparsity score feature selection method. Both clustering and classification experimental results are reported in Section 4. Section 5 gives the conclusion.

2. Related works

Assume we have a set of data $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$, where n is the number of sample points and d is the feature dimension. $x_i \in R^d$ is the ith sample. The rth feature of the ith sample x_i is denoted as f_{ri}, $r = 1, \ldots, d$. Let $\mu_r = \frac{1}{n}\sum_{i=1}^{n} f_{ri}$ represent the mean of the rth feature. If there is supervised information with the data, class labels of the data are given in $\{1, 2, \ldots, l\}$, where l is the number of categories. f_r^l is the feature vector containing the rth feature, which belongs to the lth class. Let μ_r^l denote the mean of class l corresponding to the rth feature. n_l denotes the number of data points belonging to the lth category.

Laplacian Score is obtained to reflect the locality preserving power of each feature, which is based on the assumption that, two data points from the same class should be close to each other. Laplacian Score seeks those features that respect the local geometric structure via a nearest neighbor graph. Let LS_r denote the Laplacian Score of the rth feature, which should be minimized as follows [21]:

$$\mathrm{LS}_r = \frac{\sum_{i,j}(f_{ri} - f_{rj})^2 S_{ij}}{\sum_i (f_{ri} - \mu_r)^2 D_{ii}} \tag{1}$$

where D is a diagonal matrix and $D_{ii} = \sum_j S_{ij}$, and S_{ij} is defined by the local relationship between sample points x_i and x_j as follows:

$$S_{ij} = \begin{cases} e^{-\frac{||x_i - x_j||^2}{t}}, & \text{if } x_i \text{ and } x_j \text{ are close} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Here, t is a fixed parameter in experiments and the term x_i and x_j are close means that x_i is among the k nearest neighbors of x_j or x_j is among the k nearest neighbors of x_i.

Linear Discriminant Analysis (LDA) uses fisher criterion for dimensionality reduction, which also can be used for feature evaluation. Fisher Score assigns high scores to the features that can maximize the distance of data points of different classes while minimize the distance of data points of the same class. We denote the fisher score of the rth feature as FS_r and the criteria could be maximized as follows [20]:

$$\mathrm{FS}_r = \frac{\sum_{i=1}^{l} n_l (\mu_r^l - \mu_r)^2}{\sum_{i=1}^{l}\sum_{j=1}^{n_l} (f_{rj}^i - \mu_r^i)^2} \tag{3}$$

Sparsity score [22], a ℓ_1 graph-preserving feature selection method is proposed based on this graph-preserving framework. After computing the sparse reconstructive weight matrix, the sparsity score SS_r of the rth feature could be defined as follows:

$$\mathrm{SS}_r = \frac{\sum_{i=1}^{n}(f_{ri} - \sum_{j=1}^{n} \tilde{s}_{i,j} f_{rj})^2}{\frac{1}{n}\sum_{i=1}^{n}(f_{ri} - \mu_r)^2} \tag{4}$$

where $\tilde{s}_{i,j}$ is the entry of the sparse reconstruction weight matrix constructed using all data points. We will present the detail of this method in the following Section 3.3.

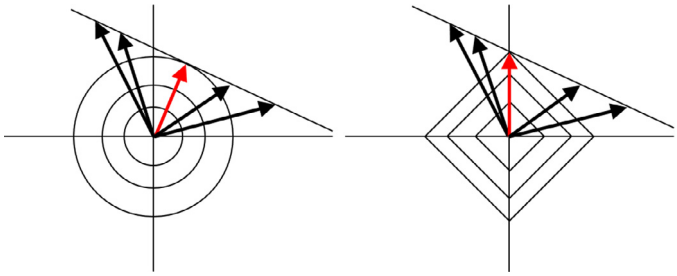


Fig. 1. Example of sparse solution.

3. Iterative sparsity score feature selection

3.1. Sparse representation

Recently, sparse representation has been extensively applied in signal processing, pattern recognition and machine learning to solve many practical problems. For instance, in signal processing fields, sparse representation is proposed as an extension signal representation such as Fourier and Wavelet representation [23,24]. Sparse representation is adopted for target recognition, signal classification and reconstruction [25–27]. Classifier based on sparse representation [28] has achieved exciting recognition rate on some face data sets. Sparse representation is related with the famous feature selection method LASSO [29]. Given a sample vector $x \in R^d$, and a matrix $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$ which contains the elements of an overcomplete dictionary in its columns. The purpose of sparse representation is to represent sample x using as fewer entries of X as possible. The objective function can be formulated as follows [30,31]:

$$\begin{aligned} \min \quad & ||s||_0 \\ s.t. \quad & x = Xs \end{aligned} \tag{5}$$

where $s \in R^n$ is the coefficient vector, and $||s||_0$ is the pseudo$-\ell_0$ norm which is the number of non-zero components in s. While ℓ_0 is not convex and to find the optimal solution of Eq. (5) is NP-hard. The above problem can be transformed to ℓ_1 and could be approximately solved by the following [28]

$$\begin{aligned} \min \quad & ||s||_1 \\ s.t. \quad & x = Xs \end{aligned} \tag{6}$$

where ℓ_1 is used to replace ℓ_0. It has been shown that the solution of ℓ_0 minimization problem is equal to the solution of ℓ_1 minimization problem under some situations [32,33]. Fig. 1 demonstrates that the ℓ_1 norm minimization can find the sparse solution, but ℓ_2 norm minimization can not get the optimal sparse solution. There have been so many works studying the equivalence of the ℓ_0 and ℓ_1 problem that the ℓ_1 approximate strategy is more reliable than other methods. The Eq. (6) can be efficiently solved via standard linear programming [31].

In practical applications, because there are always some noises in data, the constraint $x = Xs$ in Eq. (6) does not hold. Refer to [28], two robust extensions are proposed to handle this problem: (1) relax the constraint to $||x - Xs|| < \xi$, where ξ is an error tolerance; (2) replace X with $[X, I]$, where I is an d-order identity matrix. We use the first strategy in this paper.

3.2. Sparse reconstructive weights

Recently, Qiao et al. [34] constructed a sparse reconstructive weight matrix based on a modified sparse representation framework, and explained why the matrix can help to find the most compact representation of data. Assume we have n training samples $\{x_i\}_{i=1}^n$, where $x_i \in R^d$. Denote $X = [x_1, x_2, \ldots, x_n] \in R^{d\times n}$ as

data matrix. For each x_i, we can obtain the sparse reconstructive weight vector s_i through the following modified ℓ_1 minimization problem [34]:

$$\begin{aligned} \min \quad & ||s_i||_1 \\ s.t. \quad & x_i = Xs_i \\ & 1 = \mathbf{1}^T s_i \end{aligned} \tag{7}$$

where $s_i = [s_{i,1}, \ldots, s_{i,i-1}, 0, s_{i,i+1}, \ldots, s_{i,n}]^T$ is an n-dimensional vector in which the ith element is equal to zero suggesting that x_i is removed from X. The element $s_{i,\, j(i \neq j)}$ denotes the contribution of each x_j to reconstruct x_i; $\mathbf{1} \in R^n$ is a all ones vector. We find the sum-to-one constraint $1 = \mathbf{1}^T s_i$ is linear, so the modified sparse representation problem can also be solved by standard linear programming.

For each $x_i, i = 1, 2, \ldots, n$, after obtaining the corresponding weight vector $\tilde{s}_i$, the sparse reconstructive weight matrix can be defined as $S = (\tilde{s}_{ij})_{n \times n}$ as follows:

$$S = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_n]^T \tag{8}$$

where $\tilde{s}_i$ is the optimal solution of Eq. (7). It is worth noting that the weight matrix S not only reflects some intrinsic geometric properties of the data, but also preserves potential discriminant information even if no class labels are provided. The assumption is that the non-zero entries in $\tilde{s}_i$ mostly indicate the samples from the same class, so that $\tilde{s}_i$ can help to discriminate that class from the others.

As mentioned before, in many real-world problems, the constraint $x_i = Xs_i$ in Eq. (7) does not always hold. The two robust extensions in Section 3.1 can be used in the modified sparse representation to overcome the problem. The first extension is as follows:

$$\begin{aligned} \min \quad & ||s_i||_1 \\ s.t. \quad & ||x_i - Xs_i|| < \xi \\ & 1 = \mathbf{1}^T s_i \end{aligned} \tag{9}$$

where ξ is the error tolerance. As we can see, the optimal solution of Eq. (9) reflects some intrinsic geometric properties (e.g. invariant to translations and rotations) of the original data. The second extension is presented as follows:

$$\begin{aligned} min \quad & |||[s_i^T t_i^T]^T||_1 \\ s.t. \quad & \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} X & I \\ \mathbf{1}^T & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} s_i \\ t_i \end{bmatrix} \end{aligned} \tag{10}$$

where t_i is a d-dimensional vector which is induced as a reconstructive compensation term. And $\mathbf{0}$ is a d-dimensional vector of all zeros. The optimal solution of Eq. (9) is also invariant to translations, but the invariance to rotations and rescaling does not rigorously hold.

3.3. Sparsity score

Eq. (4) is the objective function of sparsity score. After several algebra operation, we can have the more compact format:

$$\begin{aligned} \mathrm{SS}_r &= \frac{\sum_{i=1}^n (f_{ri} - \sum_{j=1}^n \tilde{s}_{i,j} f_{rj})^2}{\frac{1}{n}\sum_{i=1}^n (f_{ri} - \mu_r)^2} \\ &= \frac{f_r^T (I - S - S^T + S^T S) f_r}{f_r^T (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) f_r} \end{aligned} \tag{11}$$

where S is the sparse reconstructive weight matrix via minimizing the modified ℓ_1 problem Eq. (7). As we can see the numerator in Eq. (11), if f_{ri} can be well represented by other features, the value of the numerator will be small. At the same time, variance is considered in the denominator, in which f_r with big variance can result in small value of the sparsity score. Thus the features with high values can best respect the predefined ℓ_1 graph structure (i.e. with small reconstruction error) as well as large variance.

Algorithm 1 Iterative sparsity score feature selection.

Input:
The data set X with d features, selected feature number z;
Output:
z ranked feature list
1: **repeat**
2: Construct sparse ℓ_1 graph for data X
3: Compute Sparsity Score for the features in X
4: Rank the features according to their sparsity scores in ascending order
5: Discard the last m features and update data X consisting of only the rest features
6: **until** $t \geq T$

3.4. Iterative sparsity score

Sparsity score adopts the ℓ_1 graph constructed via the whole data set, in which the features with small scores are still included. In this section, we are going to present the improved sparsity score method called iterative sparsity score. The proposed method can iteratively update the sparse ℓ_1 graph for evaluating the importance of features by its sparsity preserving ability. The key idea of iterative sparsity score is to gradually improve the sparse graph by discarding the least relevant features at each iteration.

Specifically, the algorithm finds the initial ℓ_1 graph structure from the whole dataset. After calculating the sparsity score with the initial ℓ_1 graph, we can rank the features. Then those features which respect the sparse structure are kept, while we abandon those features with small sparsity scores. In the next iterative step, the sparse ℓ_1 graph is obtained via the reserved features. Repeat the above procedure, finally the optimal structural information in the data would be unearthed.

Given data $X = [x_1, x_2, \ldots, x_n] \in R^{d \times n}$, where $x_i \in R^d$ is a d dimensional sample point. Let SS_r^t and G^t denote the sparsity score of the rth feature and the ℓ_1 graph structure in the tth iteration, respectively. We first compute the sparse ℓ_1 graph G^t by Eq. (9). Then calculate the sparsity score corresponding to the d features via Eq. (4). Discard the m features with the last m minimum sparsity score. Construct the next sparse ℓ_1 graph with the rest of the $d - m$ features. Repeat the iteration, until $t \geq T$. Here $T = \left\lfloor \frac{d-z}{m} \right\rfloor$ is the maximum iteration and z is the number of features reserved. The detail of our proposed method is shown in Algorithm 1.

It is worth noting that not only sparsity score can be extended to the iterative version, other feature selection based on score criterion also can have their iterative versions.

3.5. Multimodal iterative sparsity score

For multimodal data, a specific novel multimodal iterative sparsity score is proposed to select features across the two modalities. When we have multimodal data, it is very interesting to investigate the effectiveness of multi-modality feature selection. Conventional methods ignore the multimodal (or multi-view) information, which treat multi-modality data as traditional data. Some important information must be overlooked in the above methods. In this section, we extend the proposed feature selection method to multimodal algorithm called multimodal iterative sparsity score (MISS). Specifically, first we construct two graphs G_1 and G_2 using all the features on the two feature sets. And then in each iteratively constructing G_1^t and G_2^t process, we combine the all feature

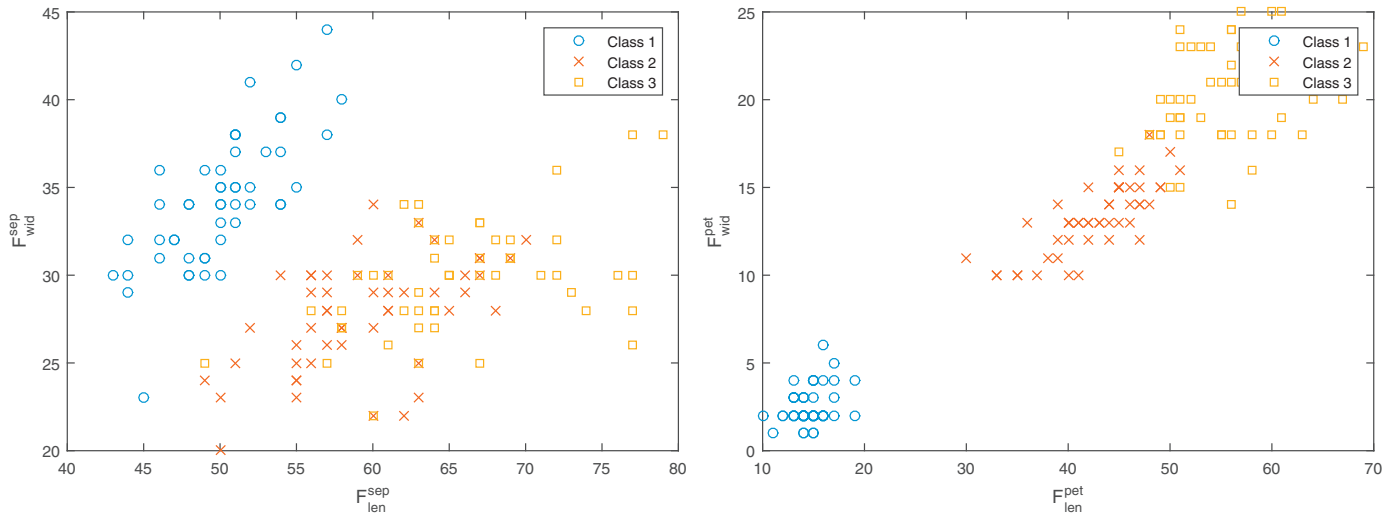


Fig. 2. Visualization of Iris data on 2 dimensions.

graph with the iterative graph as follows:

$$G_1^{t'} = G_1^t + \lambda G_2 \tag{12}$$

$$G_2^{t'} = G_2^t + \lambda G_1 \tag{13}$$

where $G_1^{t'}$ and $G_2^{t'}$ are the modified graph in each iteration and λ is a fixed parameter for balancing the contribution between these two terms. In our classification experiment, we set λ as 0.3. The modified graphs can fully fuse the information in those multi-modality data. Assume, one feature is assigned with high score in modality one, that means it is important in that modality. While in the other modality, it may be overlooked. The modified graph can relieve this situation.

4. Experiments

In this section, we carry out experiments on a lot of data sets to investigate our proposed iterative sparsity score algorithm comparing with several popular feature selection approaches. Clustering experiments are conducted on real world database. Further, we apply our proposed method on multi-modality alzheimer's disease data to verify the feature selection ability on multimodal data set.

4.1. Illustrative toy problem

In this subsection, we use iris data set from the UCI machine learning repository[2]. Iris consists of 50 instances from each of 3 categories of Iris. Four features are obtained from each sample (i.e. sepals length, sepals width, petals length and petals width). We denote the four features as F^{sep}_{len}, F^{sep}_{wid}, F^{pet}_{len} and F^{pet}_{wid}, respectively. Fig. 2 shows the distribution of Iris data on two dimensional features. As we can see from Fig. 2, class 1 can be linearly separable from the other two categories, while class 2 and class 3 are not linearly separable from each other. The Fig. 2 also indicates that feature F^{pet}_{len} and F^{pet}_{wid} contain more discriminative information since there existing large margin between class 1 and other two categories on these two features.

We compare our proposed iterative sparsity score method with other four filter-type feature selection approaches, which are variance, fisher score, Laplacian score and sparsity score. Only fisher score is supervised, while others are unsupervised. Variance ranks the features as F^{pet}_{len}, F^{sep}_{len}, F^{pet}_{wid}, F^{sep}_{len}. Fisher score sorts those features as F^{pet}_{len}, F^{pet}_{wid}, F^{sep}_{len}, F^{sep}_{wid}. When using Laplacian score with $k = 10$, the four features are sorted as F^{pet}_{wid}, F^{pet}_{len}, F^{sep}_{len}, F^{sep}_{wid}. By using sparsity score, the four features are ranked as F^{pet}_{wid}, F^{pet}_{len}, F^{sep}_{len}, F^{sep}_{wid}. Our proposed iterative sparsity score sorts those features as F^{pet}_{wid}, F^{pet}_{len}, F^{sep}_{len}, F^{sep}_{wid}.

4.2. Clustering

In this subsection, the proposed iterative sparsity score feature selection method is compared with other three unsupervised filter-type feature selection approaches (i.e. variance, Laplacian score and sparsity score) for clustering task. We do not apply fisher score, since fisher score is an supervised algorithm and there is no label information in the clustering experiments.

4.2.1. Data sets

We first simply introduce the data sets used later. The data sets we perform clustering experiments on are from UCI machine learning repository. They are wine, ionosphere, sonar, spectf heart disease, digits, and steel plate faults. Table 1 summarizes the characteristic of these UCI databases. As we can see from Table 1, the databases include binary class and multi-categories data with middle size of feature numbers.

4.2.2. Experimental settings

We first rank the features of each data set via Baseline, Variance, Laplacian Score, Sparsity Score and our proposed method. The clustering task with all the original data is denoted as Baseline. Then k-means clustering is performed on the reserved z features, where z varies from 1 to d and we can get z different clustering results. The parameter k is set as the number of categories of specific data set. This process is repeated 10 times and the average performance as well as the optimal number of selected features are reported.

F-Score metric [35] is used to measure the clustering performance, in order to compare the obtained label of each sample with that provided by the data. Assume we have a clustering result, then *Precision* and *Recall* criteria are defined as follows:

$$Precision = \frac{N_1}{N_1 + N_2} \tag{14}$$

[2] http://archive.ics.uci.edu/ml

Table 1
Characteristics of the UCI data sets.

Description	Wine	Ionosphere	Sonar	Spectf heart disease	Horse	Steel plate faults
Sample	178	351	208	267	368	1941
Feature	13	33	60	44	60	27
Class	3	2	2	2	2	7

Table 2
Clustering performance comparisons.

Dataset	Baseline	Var	LS	SS	ISS
Wine	0.5883 (13)	0.5947 (1)	0.6414 (2)	0.6631 (1)	**0.6640** (1)
Ionosphere	0.6037 (33)	0.6142 (32)	0.6404 (2)	0.6548 (1)	**0.7775** (9)
Sonar	0.5026 (60)	0.5570 (1)	0.5221 (10)	0.5649 (36)	**0.6047** (4)
Spectf heart disease	0.6618 (44)	0.6618 (43)	0.6620 (3)	0.6857 (44)	**0.6904** (27)
Horse	0.5719 (60)	0.5884 (5)	0.6536 (1)	0.6930 (1)	**0.6930** (1)
Steel plate faults	0.2561 (27)	0.2657 (20)	0.2657 (20)	0.3137 (1)	**0.3798** (19)

$$Recall = \frac{N_1}{N_1 + N_3} \tag{15}$$

where N_1 is the number of sample pairs that are clustered correctly, N_2 is the number of sample pairs which belong to the different categories, but are clustered into the same cluster, and N_3 is the number of sample pairs that belong to the same class, but are assigned to different categories. After we obtain *Precision* and *Recall*, F-Score can be calculated as follows:

$$\text{F-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

4.2.3. Experimental results

The experimental results are shown in Table 2. The best clustering results are denoted in bold font and the optimal numbers of reserved features are presented in the brackets. The parameter k in Laplacian Score is fixed as 10. The k-means clustering initialization for each repeat is the same for different algorithms for fair comparison. And the number of clusters is set as the categories of each data set.

As we can see from Table 2, the clustering performances of our proposed iterative sparsity score are usually the best one among those of all the other methods. Specifically, different feature selection methods get similar F-Score on wine data set. Because wine dataset is a quite simple database, since it contains only 13 features and 3 classes. Even that, our proposed method achieves the best performance 0.6640 with only 1 feature.

Fig. 3 demonstrates the clustering performance with different number of features on 6 UCI data sets. Iterative sparsity score methods always achieves the best clustering results with small dimension.

4.3. Classification for multimodal ADNI data

Alzheimer's disease (AD) is a physical disease and is the most common cause of dementia. In this section, we perform classification experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, in order to evaluate the effectiveness of our proposed feature selection method for classification tasks.

4.3.1. ADNI data

The data we used is from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3]. It contains 202 ADNI participants with corresponding baseline MRI and FDG-PET data. In particular, 51 AD patients, 99 MCI patients and 52 normal controls (NC) are included. In our experiment, only the data of AD patients and normal controls are used for binary classification. Fig. 4 shows MRI and PET image of a subject from ADNI database. To obtain the features, the gray matter of 93 regions of interest (ROI) are extracted as features from MR images. For PET image, we first align it to its respective MR image of the same sample using a rigid transformation. And then the average intensity of each ROI in the PET image is computed as a feature. Therefore, for each sample, we totally have 93 features from the MR image and another 93 features from the PET image.

[3] http://www.loni.usc.edu

4.3.2. Experimental settings

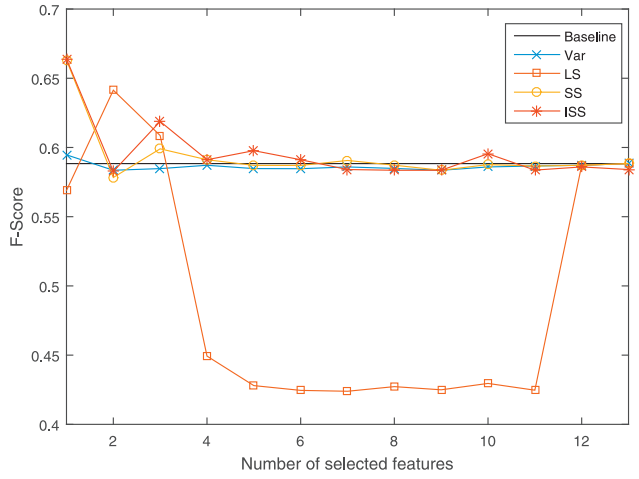
Classification performance is assessed between patients and normal controls. The proposed feature selection method is compared with 4 existing popular feature selection algorithms, including Variance, Fisher Score, Laplacian Score and Sparsity Score. The approach using all of the features for classification is denoted as Baseline. It is worth noting that except Fisher Score, other above feature selection methods are unsupervised without using the label information provided by the data.

Before classification, we adopt feature selection approach on the two modalities of ADNI data separately. For a sample now we have two sets of features from MRI and PET image, respectively. Although there are numerous feature fusion strategies, such as Canonical Correlation Analysis (CCA), Partial Least Squares (PLS) and Multi-task Learning (MTL), we use the simplest way to concatenate the two heterogeneous features into a long vector with 186 dimensions. The same choosing feature strategy in clustering is adopted for classification experiment.

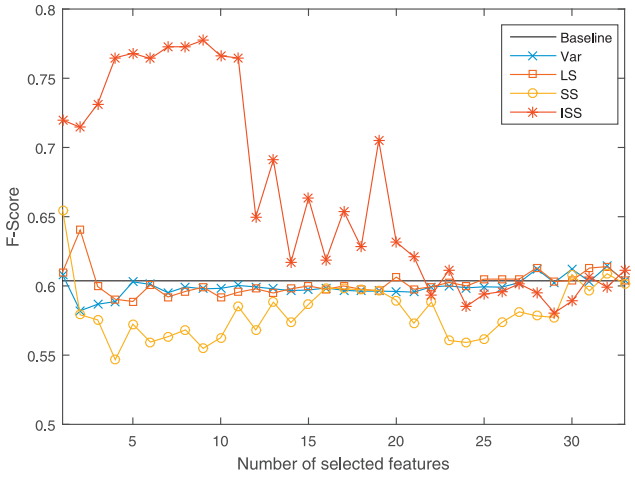
For fair comparison, we use a 10-fold cross-validation strategy to evaluate the effectiveness of our proposed method. Specifically, the whole data are equally divided into 10 subsets. For each cross-validation, the nine sample sets are used for training and the remaining data are chosen for testing. The process is independently repeated 10 times to avoid any bias introduced by randomly partitioning the dataset in cross-validation. In this experiment, we apply LIBSVM [36] for classification with a RBF kernel and default parameters. The classification accuracy is used for evaluate the performance of different methods.

4.3.3. Experimental results

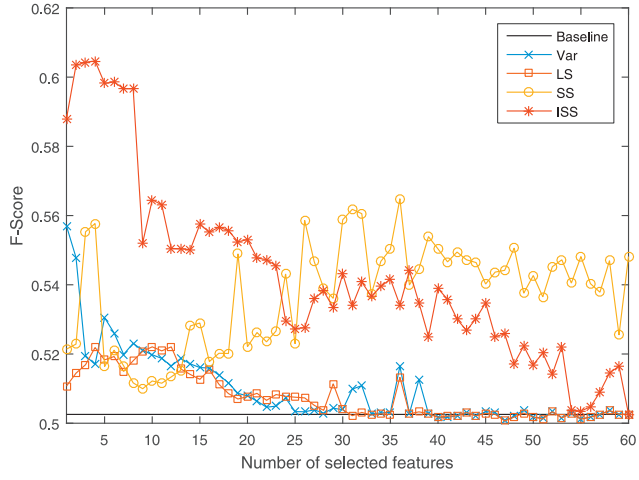
The classification accuracies are presented in Table 3 and the selected dimension corresponding to the best accuracy is shown in the brackets. From Table 3, it is shown that the classification accuracy is only 87.39% when using all the features, which is below that after feature selection. It indicates the effectiveness of feature selection. The performance of our proposed method outperforms other methods, except the supervised Fisher score. That is



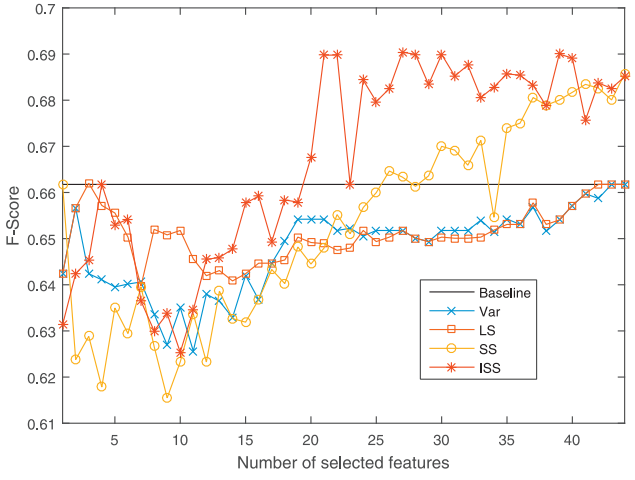
(a) Wine



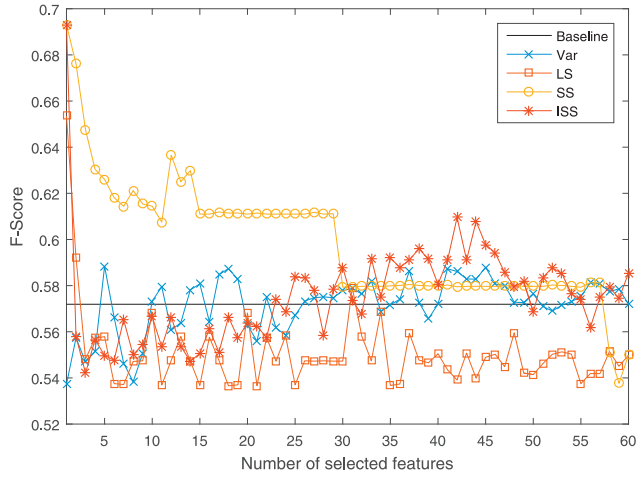
(b) Ionosphere



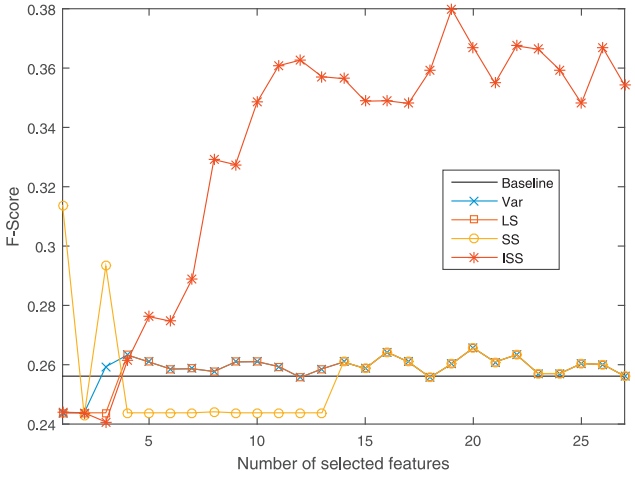
(c) Sonar



(d) Spectf heart disease



(e) Horse



(f) Steel plate faults

Fig. 3. Clustering performance with different number of features.

Table 3
Classification performance comparisons.

Method	Baseline	Var	FS	LS	SS	ISS	MISS
Accuracy	87.36 (93)	90.27 (67)	91.45 (33)	91.18 (52)	90.27 (56)	91.36 (25)	92.36 (38)

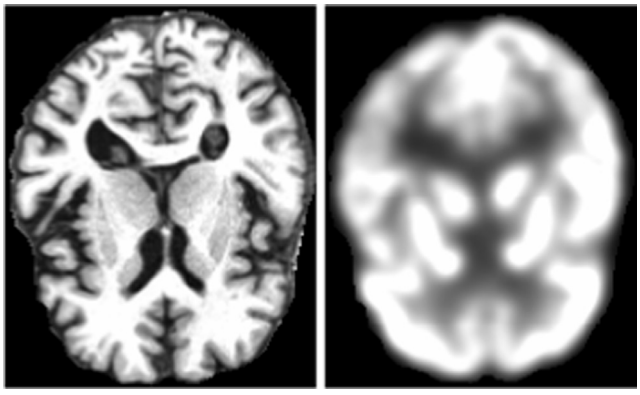


Fig. 4. Example of MRI (left) and PET (right) data.

probably because even our method is unsupervised, the structural information preserved by ℓ_1 graph is close to the real data information. While the accuracy achieved by Sparsity Score is lower than those of Fisher Score and Laplacian Score, we guess that the features with low sparsity score induce noise when constructing the sparse graph. The best classification performance is achieved by multimodal iterative sparsity score. That is because the multimodal extension of our method can fully exploit the multimodality in the data.

5. Conclusion

In this paper, a novel iterative sparsity score algorithm and its multimodal extension are proposed for feature selection. The proposed method can evaluate the importance of features in a iterative way, so that the ℓ_1 graph can be constructed more precisely Experimental results on both conventional clustering and multimodality classification demonstrate the effectiveness of the proposed methods compared with other popular filter-type feature selection algorithms. Specifically, our proposed method achieves the best clustering results on UCI datasets. Although it is unsupervised, the classification result on ADNI database outperforms other unsupervised approaches and is comparable to that of Fisher Score, which involves the label information.

In our future work, we try to embed category information into the proposed method and want to see it may get better results than Fisher Score. Constructing ℓ_1 graph in an iterative way is time-consuming especially when there are large data samples. We will design a fast algorithm to solve this problem. As we know, ℓ_2 graph also can investigate some intrinsic information from data, it is interesting to combine graphs constructed via ℓ_1 and ℓ_2 together.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61473149 and 61422204) and the NUAA Fundamental Research Funds (No. NE2013105).

References

[1] M. Liu, D. Zhang, D. Shen, View-centralized multi-atlas classification for alzheimer's disease diagnosis, Human Brain Mapp. 36 (5) (2015) 1847–1865.
[2] Y. Wang, G. Ma, L. An, F. Shi, P. Zhang, X. Wu, J. Zhou, D. Shen, Semi-supervised tripled dictionary learning for standard-dose pet image prediction using low-dose pet and multimodal mri, IEEE Trans. Biomed. Eng. (2016).
[3] J. Zhang, Y. Gao, Y. Gao, B. Munsell, D. Shen, Detecting anatomical landmarks for fast alzheimer's disease diagnosis (2016).
[4] C.O.S. Sorzano, J. Vargas, A.P. Montano, A survey of dimensionality reduction techniques, arxiv:1403.2877(2014).
[5] G. Wang, J. Ma, S. Yang, An improved boosting based on feature selection for corporate bankruptcy prediction, Expert Syst. Appl. 41 (5) (2014) 2353–2361.
[6] J. Zhang, J. Liang, H. Zhao, Local energy pattern for texture classification using self-adaptive quantization thresholds, IEEE Trans. Image Process. 22 (1) (2013) 31–42.
[7] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[8] M. Liu, D. Zhang, Pairwise constraint-guided sparse learning for feature selection, IEEE Trans. Cybernet. 46 (1) (2016) 298–310.
[9] A.K. Jain, R.P. Duin, J. Mao, Statistical pattern recognition: A review, Pattern Anal. Mach. Intell. IEEE Trans. 22 (1) (2000) 4–37.
[10] F. Wang, J. Liang, An efficient feature selection algorithm for hybrid data, Neurocomputing 193 (2016) 33–41.
[11] X. Wang, Y. Gao, Y. Cheng, A non-negative sparse semi-supervised dimensionality reduction algorithm for hyperspectral data, Neurocomputing 188 (2016) 275–283.
[12] J. Hu, H. Tang, K.C. Tan, H. Li, How the brain formulates memory: A spatio-temporal model research frontier, IEEE Comput. Intell. Mag. 11 (2) (2016) 56–68.
[13] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[14] G. Xia, H. Yan, J. Yang, Sparse preserving feature weights learning, Neurocomputing (2015).
[15] H. Shi, Y. Li, Y. Han, Q. Hu, Cluster structure preserving unsupervised feature selection for multi-view tasks, Neurocomputing 175 (2016) 686–697.
[16] A.L. Yuille, P.W. Hallinan, D.S. Cohen, Feature extraction from faces using deformable templates, Int. J. Comput. Vis. 8 (2) (1992) 99–111.
[17] J. Zhang, H. Zhao, J. Liang, Continuous rotation invariant local descriptors for texton dictionary-based texture classification, Comput. Vis. Image Underst. 117 (1) (2013) 56–75.
[18] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: ICML, 3, 2003, pp. 856–863.
[19] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[20] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[21] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Advances in Neural Information Processing Systems, 2005, pp. 507–514.
[22] M. Liu, D. Zhang, Sparsity score: A novel graph-preserving feature selection method, Int. J. Pattern Recognit. Artif. Intell. 28 (04) (2014) 1450009.
[23] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, Image Process. IEEE Trans. 15 (12) (2006) 3736–3745.
[24] J. Yang, J. Wright, T. Huang, Y. Ma, Image super-resolution as sparse representation of raw image patches, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[25] K. Huang, S. Aviyente, Sparse representation for signal classification, in: Advances in Neural Information Processing Systems, 2006, pp. 609–616.
[26] M.A. Davenport, M.F. Duarte, M.B. Wakin, J.N. Laska, D. Takhar, K.F. Kelly, R.G. Baraniuk, The smashed filter for compressive classification and target recognition, in: Electronic Imaging 2007, International Society for Optics and Photonics, 2007. 64980H–64980H
[27] Y. Wang, P. Zhang, L. An, G. Ma, J. Kang, F. Shi, X. Wu, J. Zhou, D.S. Lalush, W. Lin, et al., Predicting standard-dose pet image from low-dose pet and multimodal mr images using mapping-based sparse representation, Phys. Med. Biol. 61 (2) (2016) 791.
[28] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, Pattern Anal. Mach. Intell. IEEE Trans. 31 (2) (2009) 210–227.
[29] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodol.) (1996) 267–288.
[30] J.F. Murray, K. Kreutz-Delgado, Visual recognition and inference using dynamic overcomplete sparse learning, Neural Comput. 19 (9) (2007) 2301–2352.
[31] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.
[32] R.G. Baraniuk, Compressivesensing, IEEE Signal Process. Mag. 24 (4) (2007) 118–121.
[33] D.L. Donoho, Compressed sensing, Inf. Theory IEEE Trans. 52 (4) (2006) 1289–1306.
[34] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (1) (2010) 331–341.

[35] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using em, Mach. Learn. 39 (2–3) (2000) 103–134.
[36] C.-C. Chang, C.-J. Lin, Libsvm: Alibrary for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.

Chen Zu received the B.S. and M.S. degrees from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2010 and 2013, respectively. He is currently pursuing a Ph.D. degree at Nanjing University of Aeronautics and Astronautics. His current research interests include neuroimaging analysis, machine learning, pattern recognition, and data mining.

Linling Zhu received the B.S. and M.S. degrees from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2010 and 2013, respectively. Her current research interests include neuroimaging analysis, machine learning, pattern recognition, and data mining.

Daoqiang Zhang received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1999 and 2004, respectively. In 2004, he joined the Department of Computer Science and Engineering, NUAA, as a Lecturer, where he is currently a Professor. His current research interests include machine learning, pattern recognition, data mining, and medical image analysis. He has published over 100 scientific articles in refereed international journals such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Neuroimage, Human Brain Mapping*, and conference proceedings such as *International Joint Conferences on Artificial Intelligence, IEEE International Conference on Data Mining, and International Conference on Medical Image Computing and Computer Assisted Interventions*. Dr. Zhang is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence and the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation.