Eigenvector Sensitive Feature Selection for Spectral Clustering

Yi Jiang and Jiangtao Ren

Sun Yat-sen University, Guangzhou, 510006, P.R. China
jiangyi5@student.sysu.edu.cn, issrjt@mail.sysu.edu.cn

Abstract. Spectral clustering is one of the most popular methods for data clustering, and its performance is determined by the quality of the eigenvectors of the related graph Laplacian. Generally, graph Laplacian is constructed using the full features, which will degrade the quality of the related eigenvectors when there are a large number of noisy or irrelevant features in datasets. To solve this problem, we propose a novel unsupervised feature selection method inspired by perturbation analysis theory, which discusses the relationship between the perturbation of the eigenvectors of a matrix and its elements' perturbation. We evaluate the importance of each feature based on the average $L1$ norm of the perturbation of the first k eigenvectors of graph Laplacian corresponding to the k smallest positive eigenvalues, with respect to the feature's perturbation. Extensive experiments on several high-dimensional multi-class datasets demonstrate the good performance of our method compared with some state-of-the-art unsupervised feature selection methods.

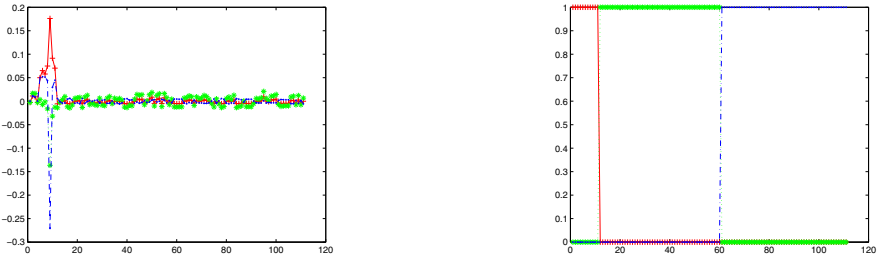
Keywords: Feature Selection, Graph Laplacian, Perturbation Analysis.

1 Introduction

Spectral clustering has wide applications ranging from text, image, web, bioinformatics to social science, for exploratory data analysis. Roughly speaking, spectral clustering is the technique to partition the rows of a matrix into multiple clusters based on the few top eigenvectors of graph Laplacian[9]. Compared with classical methods like k-means and mixture models, it has three advantages. Firstly, it doesn't need any explicit or implicit assumptions about the sample distribution. Secondly, it is easy to implement and has polynomial time solutions. Lastly, it is equivalent to graph cut problems, which are well developed. Due to these virtues, there are enormous literatures in the past on spectral clustering[6]-[21], but the nature of spectral clustering remains unchanged: *The performance of spectral clustering is determined by the quality of the chosen eigenvectors of graph Laplacian.*

However, recently, Tao Xiang, etc[10] pointed out that the first k eigenvectors of graph Laplacian may be uninformative and inappropriate for spectral clustering given noisy, irrelevant and high-dimensional data. Note that '*the first k eigenvectors*' denotes the eigenvectors corresponding to the k smallest positive eigenvalues, and '*the first k eigenvector*' denotes the eigenvector with the

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 114–129, 2011.
© Springer-Verlag Berlin Heidelberg 2011



(a) Soft cluster indicators (b) Ideal cluster indicators

Fig. 1. The distribution of the soft and ideal cluster indicators for CLL-SUB-111

k smallest positive eigenvalue[9]. For the demonstration of the impact of irrelevant features on graph Laplacian's eigenvectors, we provide an intuitive example with a dataset *CLL-SUB-111*[1] which has 3 classes and 11340 features. In Fig.1, each curve in (a) represents the distribution of the components of one of the first three eigenvectors of its graph Laplacian computed with its full 11340 features, and the curve of the same color in (b) is the 'ideal' distribution. It is clear that each distribution in (a) has only one peak region between 0 and 20, suggesting that spectral clustering will group these samples into two clusters based on the inappropriate graph Laplacian, which differs from the 'true' cluster structure. This example demonstrates that the graph Laplacian constructed from the full features may degrade the performance of spectral clustering when there are a large number of irrelevant and noisy features in the high-dimensional dataset, hence we need to perform feature selection before constructing the graph Laplacian for spectral clustering.

The core problem of feature selection is how to evaluate the importance of features, which has numerous criterions such as Laplacian Score(LS)[36], Spec[37], MCFS[39], FSFS[34], FCBF[35], FSSEM[30] and EVSC[41], etc. In the recent development of spectral clustering, Ling Huang, etc[11]-[13] present some proofs of the close relationship between the perturbation of clustering result and laplacian graph's eigenvectors due to the perturbation of data. These researches inspire us that the perturbation of the feature values of data will have impact on the perturbation of the eigenvectors of graph Laplacian and the result of spectral clustering, hence we can evaluate the importance of features by using the perturbation of the eigenvectors of graph Laplacian in respect of the perturbation of each feature.

In this paper, we propose a new feature evaluation criterion based on the recent developments of perturbation analysis[2][11]-[15]. Specifically, to evaluate a feature's importance, we perturb the value of this feature by introducing a perturbation factor to it for all the samples in the data set. This will induce a perturbation of the similarity matrix, and in turn a perturbation of the graph Laplacian. Finally, this leads to the perturbation of the eigenvectors of graph Laplacian. It is natural to believe that if a small perturbation of one feature

[1] `http://featureselection.asu.edu/datasets.php`

induces a great perturbation of the eigenvectors of graph laplacian, this feature is important for spectral clustering. Then, we use the average L1-norm of the perturbation of the first k eigenvectors of graph Laplacian in terms of the small perturbation of one feature to estimate the significance of this feature. The criterion is referred to as *Eigen Vector Sensitive Feature Selection Criterion (EVSFSC)*. Based on this criterion, we can perform feature selection for spectral clustering. Extensive experiment results over six real-world datasets demonstrate the superiority of our method compared with four traditional unsupervised feature selection methods.

2 Feature Selection Based on Perturbation Analysis

In this section, we study the perturbation of graph Laplacian's eigenvectors in terms of the perturbation of each feature, with three different definitions of graph Laplacian L, L_{rw} and L_{sym}[9]. Based on these analysis, we formulate three feature evaluation criterions, then a feature selection algorithm is proposed for the most common spectral clustering algorithms.

2.1 Problem Definition

For a dataset $X = \{x^i\}_{i=1}^n$, $x^i \in R^{K\times 1}$ represents the i-th data sample, where K is the dimension of X, and x_t^i denotes the t-th feature value of x^i. Suppose S, D and L are similarity matrix, diagonal degree matrix and graph Laplacian respectively, $S_{i,j}$ represents the similarity between x^i and x^j, $D = diag(S\mathbf{1})$ ($\mathbf{1} = (1, ..., 1)^T$) and $L = D - S$.

Let ξ be a perturbation factor, if we perturb X on the t-th feature with ξ, which means $\hat{x}_t^i = x_t^i + \xi x_t^i$, $i = 1, ..., n$, and keep other features unchanged, then we get a perturbed dataset $\hat{X}_t = \{\hat{x}^i\}_{i=1}^n$. Let $\hat{L}_t$ be the perturbed graph Laplacian based on $\hat{X}_t$. Suppose $\hat{q}_{t,r}$ and q_r are the r-th eigenvector of $\hat{L}_t$ and L respectively, then the perturbation of the r-th eigenvector of graph Laplacian L caused by the perturbation of the t-th feature can be defined as $\triangle q_{t,r} = \hat{q}_{t,r} - q_r$. The greater the L1 norm of $\triangle q_{t,r}$ is, the more important the t-th feature is. Thus, our main problem is how to evaluate $\triangle q_{t,r}$ with respect to ξ. Let's begin by proving the relationship between $\hat{D}_t, \hat{L}_t$ and D, L, where $\hat{D}_t$ is the perturbed similarity matrix based on $\hat{X}_t$.

In this paper, we adopt RBF function as the similarity measure between data samples, and our framework can also be easily extended to other popular similarity measures such as dot product, square Euclidean, etc. Then $S_{i,j}$ can be formulated as

$$S_{i,j} = e^{-\frac{\sum_{h=1}^{K}(x_h^i - x_h^j)^2}{2\delta^2}}$$

where δ^2 is the kernel bandwidth. When we perturb the t-th feature with factor ξ, which means $\hat{x}_t^i = (1+\xi)x_t^i$, $i = 1, ..., n$, the perturbed similarity $\hat{S}_{t,i,j}$ is

$$\hat{S}_{t,i,j} = e^{-\frac{\sum_{h=1,h\neq t}^{K}(x_h^i - x_h^j)^2 + (1+\xi)^2(x_t^i - x_t^j)^2}{2\delta^2}} \tag{1}$$

Now we can derive the relationship between $\hat{D}_t, \hat{L}_t$ and D, L as follows.

Theorem 1. *If $\xi \to 0$, $\hat{D}_t$ and $\hat{L}_t$ can be approximated by*

$$\hat{D}_t \approx D - \xi D_t^1, \quad \hat{L}_t \approx L - \xi L_t^1$$

Then,

$$\hat{D}_t - D \approx -\xi D_t^1, \quad \hat{L}_t - L \approx -\xi L_t^1 \tag{2}$$

where $S_{t,i,j}^1 = S_{i,j}\frac{(x_t^i - x_t^j)^2}{\delta^2}$, $D_{t,i,i}^1 = \sum_{h=1}^n S_{t,i,h}^1$, $L_t^1 = D_t^1 - S_t^1$, D_t^1 is a diagonal matrix.

Proof. based on formula (1), we can get

$$\frac{\partial \hat{S}_{t,i,j}}{\partial \xi} = -(\xi + 1)\hat{S}_{t,i,j}\frac{(x_t^i - x_t^j)^2}{\delta^2}$$

Then, when $\xi \to 0$, we can derive the first-order Taylor expansion for $\hat{S}_{t,i,j}$

$$\hat{S}_{t,i,j} = S_{i,j} - S_{i,j}\frac{(x_t^i - x_t^j)^2}{\delta^2} \cdot \xi + O(\xi)$$

and we can get

$$\hat{S}_{t,i,j} - S_{i,j} \approx -S_{i,j}\frac{(x_t^i - x_t^j)^2}{\delta^2} \cdot \xi \approx -\xi \cdot S_{t,i,j}^1$$

$$\hat{D}_{t,i,i} = \sum_{h=1}^n \hat{S}_{t,i,h} \approx \sum_{h=1}^n S_{i,h} - \xi \cdot \sum_{h=1}^n S_{t,i,h}^1 \approx D_{i,i} - \xi \cdot D_{t,i,i}^1$$

Thus,

$$\hat{D}_t - D \approx -\xi D_t^1$$

$$\hat{L}_t - L = (\hat{D}_t - D) - (\hat{S}_t - S) = -\xi \cdot (\hat{D}_t^1 - \hat{S}_t^1) \approx -\xi L_t^1$$

□

In general, $L = D - S$ is the unnormalized graph Laplacian[9]. Moreover, there are two other normalized graph Laplacians[9] $L_{rw} = D^{-1}L$ and $L_{sym} = D^{-1/2}LD^{-1/2}$. We will derive $\triangle q_{t,r}$, $\triangle q_{rw,t,r}$ and $\triangle q_{sym,t,r}$ with respect to L, L_{rw} and L_{sym} respectively in the following sections.

2.2 $\triangle q_{t,r}$ with Respect to L

Perturbation analysis theory [2] discusses the relationship between the perturbation of the eigenvectors of a matrix and its elements' perturbation, which will be summarized in **Theorem 2**.

Theorem 2. *(*First-Order Eigenvector Perturbation*)*
Let A and B be matrices with elements which satisfy the relations: $|A_{ij}| < 1$ and $|B_{ij}| < 1$, and A has the normalized eigenvector set $\{q_r\}_{r=1}^n$ and eigenvalue set $\{\lambda_r\}_{r=1}^n$, where the multiplicity of any eigenvalue is 1, if $\xi \to 0$, then the r-th eigenvector $\hat{q}_r$ of $A + \xi B$ is approximately expressed as:

$$\hat{q}_r \approx q_r + \xi \cdot \{\sum_{h=1, h\neq r}^n \frac{q_h^T B q_r}{\lambda_r - \lambda_h} q_h\}. \tag{3}$$

Based on **Theorem 1** and **Theorem 2**, we can easily derive $\triangle q_{t,r} = \hat{q}_{t,r} - q_r$, which is summarized in **Theorem 3**. It is worth pointing out that the conditions $|A_{ij}| < 1$ and $|B_{ij}| < 1$ can be satisfied for RBF kernel.

Theorem 3. *Let* $\{\lambda_r\}_{r=1}^n$ *and* $\{q_r\}_{r=1}^n$ *be the eigenvalue and normalized eigenvector sets of* $Lq_r = \lambda_r q_r$, *and* $\lambda_1 < \lambda_2 < \ldots < \lambda_n$, *if* $\xi \to 0$, *then the r-th eigenvector of* $\hat{L}_t$ *based on* $\hat{X}_t$ *can be approximated by*

$$\hat{q}_{t,r} \approx q_r + \xi \cdot p_{t,r}$$

Then,

$$\triangle q_{t,r} = \hat{q}_{t,r} - q_r \approx \xi \cdot p_{t,r} \tag{4}$$

where

$$p_{t,r} = -\sum_{h=1, h\neq r}^{n} \Big(\frac{q_h^T L_t^1 q_r}{\lambda_r - \lambda_h} \Big) q_h$$

Proof. this can be proved with **Theorem 1** and **Theorem 2**. □

2.3 $\triangle q_{rw,t,r}$ with Respect to L_{rw}

For computing the r-th eigenvector's perturbation $\triangle q_{rw,t,r} = \hat{q}_{rw,t,r} - q_{rw,r}$ of L_{rw}, where $q_{rw,r}$ is the r-th eigenvector of L_{rw} based on X, and $\hat{q}_{rw,t,r}$ is the r-th eigenvector of $\hat{L}_{rw,t}$ based on $\hat{X}_t$, we first borrow the following definition from [4], which provides some solutions for the algebraic eigenvalue problems.

Definition 1 Hermitian Definite Pencil[4]
A Hermitian definite pencil{A,B} ($A \in R^{n\times n}$ and $A \in R^{n\times n}$) is a generalized Hermitian eigenvalue problem: $Aq = \lambda Bq$, where A and B are Hermitian, that is, **if the conjugate transpose of matrix A or B is denoted by** A^* **or** B^*, then $A^* = A$ and $B^* = B$, and A or B or $\alpha A + \beta B$ for some scalars α and β is positive definite, q and λ are the corresponding eigenvector and eigenvalue respectively.

Since $L = L^*$ and $\forall$ x, $x^T Dx > 0$, then $\{L, D\}$ is a Hermitian definite pencil. For the proof of **Theorem 4**, we describe one property for L_{rw} and two properties for $\{L, D\}$ in **Property 1**, which can be found in [9] and [3] respectively.

Property 1
(a) The eigen-system of L_{rw} is equal to that of the Hermitian definite pencil $\{L, D\}$. That is, $\forall\ r \in \{1, ..., n\}$, $L_{rw} q_{rw,r} = \lambda_{rw,r} q_{rw,r} \Leftrightarrow Lq_{rw,r} = \lambda_{rw,r} D q_{rw,r}$

(b) For $\{L, D\}$, if $\lambda_{rw,r} \neq \lambda_{rw,r1}$, $q_{rw,r}^T D q_{rw,r1} = 0$, and if $q_{rw,r}$ is a normalized eigenvector, then $q_{rw,r}^T D q_{rw,r} = 1$.

(c) If $\{L, D\}$ has the eigenvalue and eigenvector sets $\{\lambda_{rw,r}\}_{r=1}^n$ and $\{q_{rw,r}\}_{r=1}^n$, and the multiplicity of any eigenvalue is 1, then $\{q_{rw,r}\}_{r=1}^n$ constitute a basis for R^n.

Then we can propose **Theorem 4** for $\triangle q_{rw,t,r}$ in the following.

Theorem 4. *For* $\{L,D\}$, *let* $\{\lambda_{rw,r}\}_{r=1}^n$ *and* $\{q_{rw,r}\}_{r=1}^n$ *be the corresponding eigenvalue and normalized eigenvector sets, and* $\lambda_{rw,1} < \lambda_{rw,2} < \ldots < \lambda_{rw,n}$, *if*

$\xi \to 0$, the r-th perturbed eigenvector $\hat{q}_{rw,t,r}$ of the perturbed normalized graph Laplacian $\hat{L}_{rw,t}$ based on $\hat{X}_t$ can be approximated as

$$\hat{q}_{rw,t,r} = q_{rw,r} + \xi \cdot p_{rw,t,r} + O(\xi \cdot \mathbf{1})$$

Then,

$$\triangle \mathbf{q_{rw,t,r}} = \hat{\mathbf{q}}_{\mathbf{rw,t,r}} - \mathbf{q_{rw,r}} = \xi \cdot \mathbf{p_{rw,t,r}} + \mathbf{O}(\xi \cdot \mathbf{1}) \tag{5}$$

where $p_{rw,t,r} = \left\{ \sum_{h=1, h\neq r}^{n} \left(\frac{q_{rw,h}^T \{\lambda_{rw,r} D_t^1 - L_t^1\} q_{rw,r}}{\lambda_{rw,r} - \lambda_{rw,h}} \right) q_{rw,h} + \left(\frac{q_{rw,r}^T D_t^1 q_{rw,r}}{2} \right) q_{rw,r} \right\}$.

Proof. Based on **Theorem 1**, if $\xi \to 0$, then

$$\hat{D}_t - D = -\xi \cdot D_t^1 + O(\xi \cdot \mathbf{I}) \quad and \quad \hat{L}_t - L = -\xi \cdot L_t^1 + O\{\xi \cdot (\mathbf{1}^T \cdot \mathbf{1})\}.$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1} = (1, ..., 1)^T$.
It is natural that[2]

$$\hat{\lambda}_{rw,t,r} - \lambda_{rw,r} = \xi \cdot \eta_{rw,t,r} + O(\xi \cdot \mathbf{1}) \tag{6}$$

$$\hat{q}_{rw,t,r} - q_{rw,r} = \xi \cdot p_{rw,t,r} + O(\xi \cdot \mathbf{1}) \tag{7}$$

Now our goal is to estimate the column vector $p_{rw,t,r}$.
Because of **Property 1 (a)**, we get

$$\hat{L}_t \hat{q}_{rw,t} = \hat{\lambda}_{rw,t} \hat{D}_t \hat{q}_{rw,t} \tag{8}$$

Based on formula (2) and (6)-(8), we get

$$\{L - \xi \cdot L_t^1\}\{q_{rw,r} + \xi \cdot p_{rw,t,r}\} = \{\lambda_{rw,r} + \xi \cdot \eta_{rw,t,r}\}\{D - \xi \cdot D_t^1\}\{q_{rw,r} + \xi \cdot p_{rw,t,r}\} \tag{9}$$

When $\xi \to 0$, (9) can be rewritten as

$$L p_{rw,t,r} - L_t^1 q_{rw,r} = -\lambda_{rw,r} D_t^1 q_{rw,r} + \lambda_{rw,r} D p_{rw,t,r} + \eta_{rw,t,r} D q_{rw,r} \tag{10}$$

With **Property 1**(c), $p_{rw,t,r}$ can be expressed as a linear combination of $\{q_{rw,r}\}_{r=1}^n$, that is,

$$p_{rw,t,r} = \sum_{h=1}^{n} \varepsilon_{r,h} q_{rw,h} \tag{11}$$

Substitute (11) into (10), and left multiply (10) by $q_{rw,r1}^T (r1 \neq r)$, we get

$$\begin{aligned} \sum_{h=1}^{n} \lambda_{rw,h} \varepsilon_{r,h} q_{rw,r1}^T D q_{rw,h} - q_{rw,r1}^T L_t^1 q_{rw,r} = -\lambda_{rw,r} q_{rw,r1}^T D_t^1 q_{rw,r} + \\ \lambda_{rw,r} \sum_{h=1}^{n} \varepsilon_{r,h} q_{rw,r1}^T D q_{rw,h} + \eta_{rw,t,r} q_{rw,r1}^T D q_{rw,r} \end{aligned} \tag{12}$$

With **Property 1** (b), we get

$$\varepsilon_{r,r1} = \frac{q_{rw,r1}^T\{\lambda_{rw,r}D_t^1 - L_t^1\}q_{rw,r}}{\lambda_{rw,r} - \lambda_{rw,r1}} \tag{13}$$

For $\{\hat{L}_t, \hat{D}_t\}$, which is also a Hermitian definite pencil,

$$\{q_{rw,r}^T + \xi p_{rw,t,r}^T\}\{D - \xi D_t^1\}\{q_{rw,r} + \xi p_{rw,t,r}\} = 1 \tag{14}$$

With $\xi \to 0$ and (11), (14) can be rewritten as

$$\varepsilon_{rr} = \frac{q_{rw,r}^T D_t^1 q_{rw,r}}{2} \tag{15}$$

Finally, based on (13) and (15), $p_{rw,t,r}$ can be computed by

$$p_{rw,t,r} = \Big\{ \sum_{h=1,h\neq r}^{n} \Big(\frac{q_{rw,h}^T\{\lambda_{rw,r}D_t^1 - L_t^1\}q_{rw,r}}{\lambda_{rw,r} - \lambda_{rw,h}}\Big)q_{rw,h} + \Big(\frac{q_{rw,r}^T D_t^1 q_{rw,r}}{2}\Big)q_{rw,r}\Big\}$$

□

2.4 $\triangle q_{sym,t,r}$ with Respect to L_{sym}

The spectral clustering theories [9] reveal that if $q_{rw,r}$ is the eigenvector of L_{rw} with $\lambda_{rw,r}$, then $q_{sym,r} = D^{1/2}q_{rw,r}$ is the eigenvector of L_{sym} with the same eigenvalue. Based on this connection between the eigen-system of L_{rw} and L_{sym}, and **Theorem 4**, the calculation of $\triangle q_{sym,t,r}$ for L_{sym} is shown in **Theorem 5**.

Theorem 5. *With the conditions of **Theorem 4**, let $q_{sym,r}$ be the normalized eigenvector of L_{sym}, and $\hat{q}_{sym,t,r}$ be the corresponding r-th eigenvector of $\hat{L}_{sym,t}$ based on $\hat{X}_t$, if $\xi \to 0$, then $\hat{q}_{sym,t,r}$ can be approximated as*

$$\hat{q}_{sym,t,r} = q_{sym,r} + \xi \cdot p_{sym,t,r} + O(\xi \cdot \mathbf{1})$$

Then,

$$\triangle \mathbf{q_{sym,t,r}} = \hat{\mathbf{q}}_{\mathbf{sym,t,r}} - \mathbf{q_{sym,r}} = \xi \cdot \mathbf{p_{sym,t,r}} + \mathbf{O}(\xi \cdot \mathbf{1}) \tag{16}$$

where $p_{sym,t,r} = \Big\{ -\frac{1}{2}D^{-1/2}D_t^1 q_{rw,r} + D^{1/2}p_{rw,t,r}\Big\}$

Proof. If $\xi \to 0$, then

$$\hat{D}_t^{1/2} = \{D - \xi \cdot D_t^1 + O(\xi \cdot \mathbf{I})\}^{1/2} \approx D^{1/2}(\mathbf{I} - \xi \cdot D^{-1}D_t^1)^{1/2} \tag{17}$$

where $\mathbf{I}$ is the identity matrix.
Then, the first order Taylor expansion of $\hat{D}_t^{1/2}$ can be rewritten as

$$(\mathbf{I} - \xi \cdot D^{-1}D_t^1)^{1/2} \approx \mathbf{I} - \frac{\xi}{2} \cdot D^{-1}D_t^1 + O(\xi \cdot \mathbf{I}) \tag{18}$$

Based on formula (17) and (18), we get

$$\hat{D}_t^{1/2} = D^{1/2} - \frac{\xi}{2} \cdot D^{-1/2} D_t^1 + O(\xi \cdot \mathbf{I})$$

Thus

$$\hat{q}_{sym,t,r} = \hat{D}_t^{1/2} \hat{q}_{rw,t,r} = q_{sym,r} + \xi \cdot \Big\{ -\frac{1}{2} D^{-1/2} D_t^1 q_{rw,r} + D^{1/2} p_{rw,t,r} \Big\} + O(\xi \cdot \mathbf{1})$$

□

2.5 Eigenvector Sensitive Feature Selection

Based on the discussion of section 2.2, 2.3 and 2.4, when the value of ξ is sufficiently small, the j-th component of the eigenvector perturbation $\triangle q_{t,r}$($\triangle q_{rw,t,r}$ or $\triangle q_{sym,t,r}$) of graph Laplacian L(L_{rw} or L_{sym}) is approximately linear with ξ, and the corresponding gradient is just the j-th component of $p_{t,r}$($p_{rw,t,r}$ or $p_{sym,t,r}$), which reflects the rate of the change of the j-th component of the r-th eigenvector of L(L_{rw} or L_{sym}) in response to the perturbation of the t-th feature. Hence, it is natural to use the $L1$ norm of $p_{t,r}$, $p_{rw,t,r}$ and $p_{sym,t,r}$ to evaluate the importance of the t-th feature to the r-th eigenvector of L, L_{rw} and L_{sym} respectively.

However, since the result of spectral clustering is determined by the first k eigenvectors of graph Laplacian, we should evaluate the importance of the r-th feature to the spectral clustering based on its impact on the first k eigenvectors of the corresponding graph Laplacian. Thus, we propose to employ the average $L1$ norm of $p_{t,r}$($p_{rw,t,r}$ or $p_{sym,t,r}$) over the first k eigenvectors of L(L_{rw} or L_{sym}) to estimate the importance of the t-th feature in the corresponding spectral clustering. This criterion is called *EigenVector Sensitive Feature Selection Criterion*(***EVSFSC***), whose formal definitions are expressed as follows.

When the graph Laplacian is L, for the t-th feature, then

$$EVSFSC(t) = \frac{1}{k} \sum_{r=2}^{k+1} \|p_{t,r}\|_1 \tag{19}$$

Similarly, when the graph Laplacian is L_{rw}, for the t-th feature, then

$$EVSFSC(t) = \frac{1}{k} \sum_{r=2}^{k+1} \|p_{rw,t,r}\|_1 \tag{20}$$

Finally, when the graph Laplacian is L_{sym}, for the t-th feature, then

$$EVSFSC(t) = \frac{1}{k} \sum_{r=2}^{k+1} \Big\| p_{sym,t,r} \|_1 \tag{21}$$

Algorithm 1. Eigenvector Sensitive Feature Selection for Spectral Clustering

Input: Data set **X**, Feature number **m**, Spectral clustering type **SCT**
Output: Feature subset F_m
1. Construct the similarity matrix S with RBF function
2. Build L and D based on S
3. If **SCT** $==' \mathbf{USC}'$
4. Calculate the eigen-system (λ_r, q_r) of L, $1 \leq r \leq n$.
5. Else if **SCT**==**'NSCLrm'**
6. Calculate the eigen-system (λ_r, q_r) of $L_{rw} = D^{-1}L$, $1 \leq r \leq n$,
7. Else
8. Calculate the eigen-system (λ_r, q_r) of $L_{sym} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$, $1 \leq r \leq n$.
9. End if
10. Normalize the eigenvectors of $\{q_r\}_{r=1}^{n}$.
for $t = 1$ **to** K **do**
11. Calculate the EVSFSC of the t-th feature based on (19), (20) or (21).
end for
12. Rank the features decreasingly according to the value of EVSFSC and select the leading m features, that is $F_m = \{F_{K_1}, ..., F_{K_m}\}$
13. return F_m

2.6 Eigenvector Sensitive Feature Selection for Spectral Clustering

Based on the criterions of (19)-(21), we summarize the eigenvector sensitive feature selection for spectral clustering algorithm in **Algorithm 1**. In this algorithm, **SCT** represents the type of spectral clustering, **USP** represents the unnormalized spectral clustering, **NSCLrm** and **NSCLsym** represent the normalized spectral clustering with L_{rm} and L_{sym} respectively. The computation complexity for main steps is listed below.

• In step 1 and 2, we need $O(n^2K)$ operations to build S, D and L;
• In step 4, 6 or 8, we need $O(n^3)$ operations to get the eigenvalues and eigenvectors of graph Laplacian by Lanczos algorithm[5];
• In step 10, we need $O(n^3K)$ operations to calculate the EVSFSC score for all features;
• In step 11, the top m features can be found within O(KlogK).
Thus, the computation complexity of **Algorithm 1** is $MAX(n^3K, KlogK)$.

3 Related Work

Spectral Clustering. The spectral clustering based on the graph cut theory is to find the best cuts of a graph according to certain predefined criterion functions such as RatioCut[6] and normalized cut(Ncut)[7]. The relaxing RatioCut leads to the unnormalized spectral clustering[9] based on the eigenvectors of unnormalized graph Laplacian $L = D - S$, while the relaxing Ncut leads to the normalized spectral clustering[7][8] based on the eigenvectors of $L_{rw} = D^{-1}L$ or $L_{sym} = D^{-1/2}LD^{-1/2}$. Recently, there are several works focusing on the impact

of small errors in data or similarity matrix on spectral clustering, based on ***the perturbation analysis***[2]. [11]-[13] derive some approximate upper bounds on the errors of k-way spectral clustering with respect to the small change of data or similarity matrix($k = 2, 3, ...$). Another line of this works is to update the information of the eigen-system of graph Laplacian in the incremental spectral clustering[14][15], given a small change of similarity matrix. Besides, there are enormous literatures discussing other subjects like the incorporation of user supervision into spectral clustering[16]-[18], and the strategy of constructing graph Laplacian for spectral clustering[19]-[21], etc.

Unsupervised Feature Selection. Most of existing methods can be classified into the three categories. Methods in the first category are wrapper approaches. These include unsupervised feature selections for K-means[22]-[25], Mixture Models[26]-[32] and PCA(Principal Components Analysis)[33]. The second category measures feature similarity based some criterions, whereby redundant features are removed. [34] and [35] are the two representatives of this kind. The third category is the spectral methods. [36]-[38] perform feature selection based on certain evaluation criterions, which are the function of the eigen-system of graph Laplacian. More recently, in [39] and [40], the feature selection problems are transformed into the regression problems, which aim to find those feature vectors aligning closely to the few top eigenvectors of graph Laplacian. In our previous work [41], a eigenvalue sensitive feature selection method is proposed. But it is different from the method of this paper. The core idea of [41] is that the feature importance should be evaluated by the gradient of the eigenvalue of graph Laplacian with respect to the weight of feature. But in this paper, we introduce the perturbation analysis theory.

4 Empirical Analysis

In this section, we perform extensive experiments to demonstrate the performance of our proposed feature selection method comparing to several popular unsupervised feature selection methods. They are FSFS[34], Laplacian Score(LS)[36], Spec[37] and MCFS[39].

4.1 Dataset Decription

Six high-dimensional and multi-class datasets are selected for the experiments, which are briefly described in Table 1. All of the datasets can be found from the Feature Selection Repository[2]. For simpleness, we use **CLL**, **ORL**, **PIX**, **TOX**, **AR** and **PIE** to represent the data sets CLL-SUB-111, orlraws10P, pixraw10P, TOX-171, warpAR10P and warpPIE10P respectively.

4.2 Evaluation Criterion

In the experiments, **Clustering Accuracy(CA)**[36] is used to evaluate the performance of spectral clustering. Based on the comparison between the predefined

[2] `http://featureselection.asu.edu/datasets.php`

Table 1. Summary of six datasets

Data Set	*Instances*	*Features*	*Classes*
CLL-SUB-111	111	11340	3
orlraws10P	100	10304	10
pixraw10P	100	10000	10
TOX-171	171	5748	4
warpAR10P	130	2400	10
warpPIE10P	210	2420	10

labels **c(i)** of all samples and the obtained labels **sc(i)** by spectral clustering, **Clustering Accuracy(CA)** is formally defined as

$$\mathbf{CA} = \frac{\sum_{i=1}^{n} \delta(\mathbf{c(i)}, map(\mathbf{sc(i)}))}{n}$$

where n is the total number of data points and $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $map(\mathbf{sc(i)})$ is the permutation mapping function that maps each cluster label **sc(i)** to the equivalent label from data corpus. Here, we use the Kuhn-Munkres algorithm[1] as the mapping function.

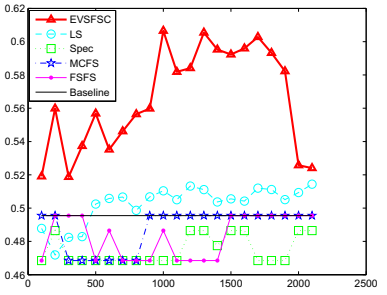
4.3 Experiment Setup

Four popular unsupervised feature selection methods are chosen as baseline methods, which are FSFS[34], Laplacian Score(LS)[36], Spec[37] and MCFS[39], and their matlab codes can be found at their homepages[3]. As discussed above, we choose **RBF function** as similarity measure, whose parameter is determined by cross-validation. Then for each dataset, the four baseline criterions and **EVSFSC** are used to select the best 100, 200,...,2100 features. Based on the selected feature subsets, the **Clustering Accuracy** of unnormalized spectral clustering with L and normalized spectral clustering with L_{rw} and L_{sym} are demonstrated in Fig.2, Fig.3 and Fig.4 respectively. And as a baseline, the **Clustering Accuracy** with the full features (without feature selection) is also depicted in all the figures, and it is referred to as 'Baseline' in the figures.

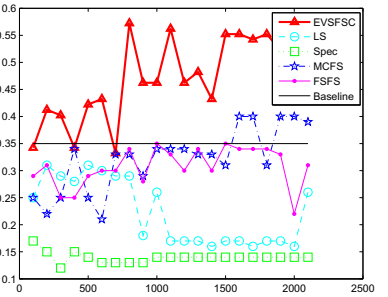
4.4 Experiment Results

Unnormalized spectral clustering with L Fig. 2(a-f) show the curves of the **Clustering Accuracy** of unnormalized spectral clustering with L versus the number of selected features on six datasets respectively, based on FSFS, Laplacian Score(LS), Spec, MCFS and EVSFSC. As we can see, our proposed algorithm achieves consistently better performance than the other methods and

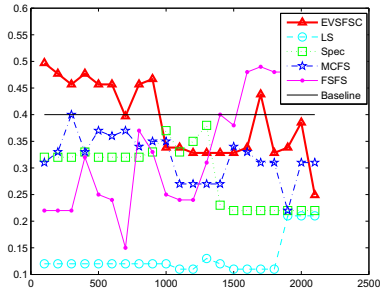
[3] `http://www.facweb.iitkgp.ernet.in/~pabitra/paper.html`, `http://www.zjucadcg.cn/dengcai/MCFS/index.html`, `http://featureselection.asu.edu/software.php`



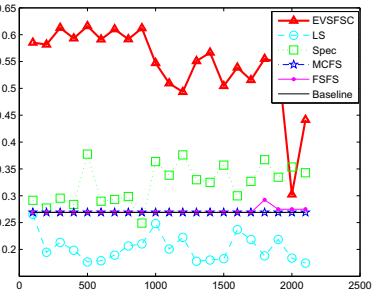
(a) CLL-SUB-111



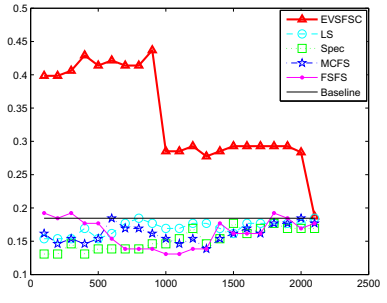
(b) Orlraws10P



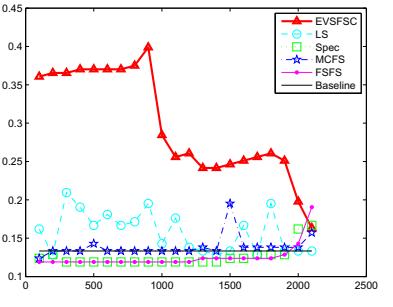
(c) Pixraw10P



(d) TOX-171



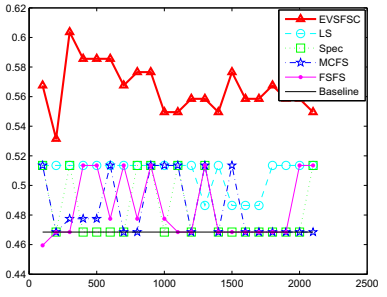
(e) WarpAR10P



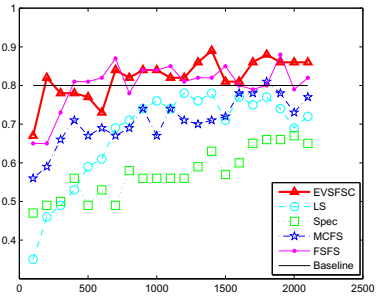
(f) WarpPIE10P

Fig. 2. Unnormalized Spectral Clustering with L

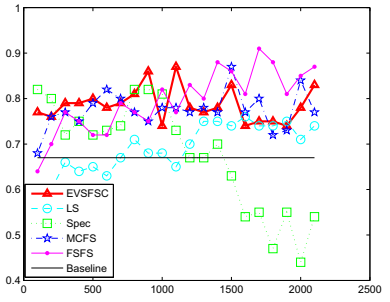
the baseline method without feature selection. Although the unnormalized spectral clustering with all features produces a poor result, most of the existing feature selection methods don't produce much better results, and sometimes produce even worse results, for example in Figure 2(b), (c) and (e). However, our method can use less than 1000 features to produce reasonably good results, whose **Clustering Accuracy** is generally higher than 0.6 on **CLL**, **ORL** and **TOX** datasets. Especially for **PIX**, **AR** and **PIE** datasets, only several hundred of selected features by our method can achieve the best results, compared with other methods.



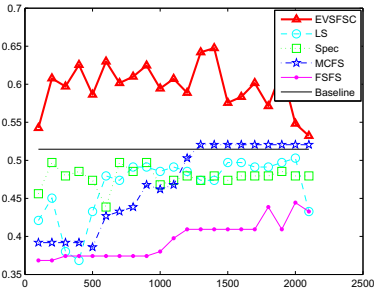
(a) CLL-SUB-111



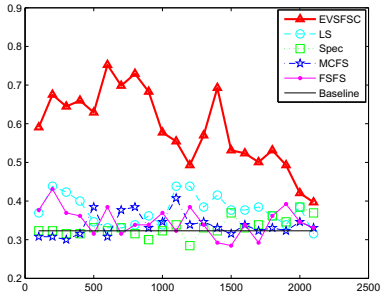
(b) Orlraw10P



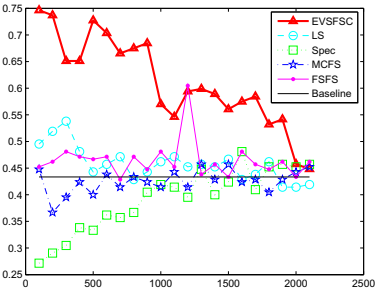
(c) Pixraw10P



(d) TOX-171



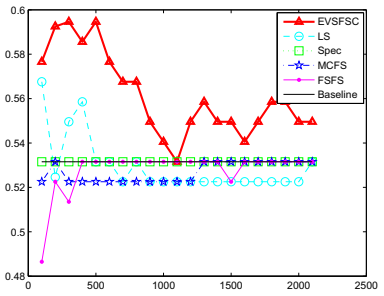
(e) WarpAR10P



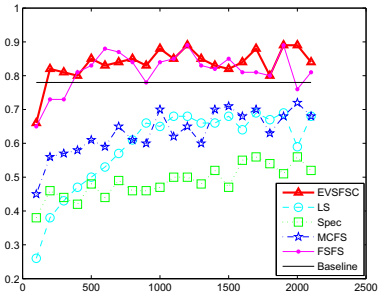
(f) WarpPIE10P

Fig. 3. Normalized Spectral Clustering with L_{rm}

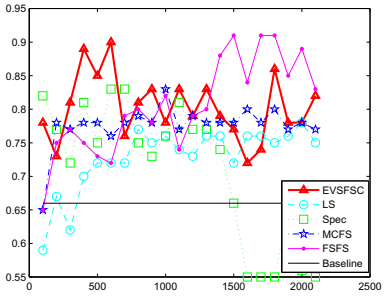
Normalized spectral clustering with L_{rw} Fig. 3(a-f) reveal the curves of the **Clustering Accuracy** of normalized spectral clustering with L_{rw} versus the number of selected features on six data sets respectively, based on **EVSFSC** and other four methods. For all of the six data sets, our method also can achieve best performance than the others. Specifically, the difference between 'Baseline' and FSFS, Laplacian Score(LS), Spec, MCFS is not obvious on **CLL**, **TOX**, **AR** and **PIE** datasets, but **EVSFSC** can still achieve great improvements.



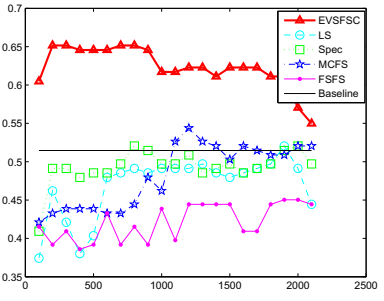
(a) CLL-SUB-111



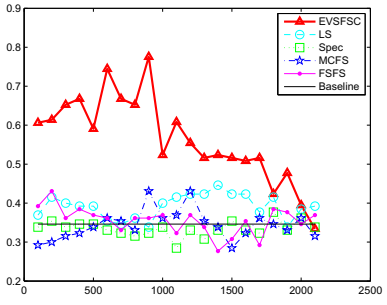
(b) Orlraws10P



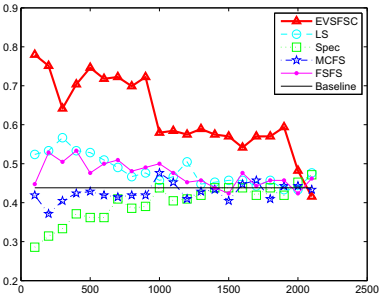
(c) Pixraw10P



(d) TOX-171



(e) WarpAR10P



(f) WarpPIE10P

Fig. 4. Normalized Spectral Clustering with L_{sym}

Normalized spectral clustering with L_{sym} Fig. 4(a-f) demonstrate the curves of the **Clustering Accuracy** of Normalized spectral clustering algorithm with L_{sym} versus the number of selected features on six data sets respectively, based on **EVSFSC** and other four methods mentioned before. Except for datasets **ORL** and **PIX**, our method significantly outperforms the other four methods. On data sets **ORL** and **PIX**, there exist some methods such as FSFS and MCFS performing comparably to our method with the increase of feature number, but EVSFSC can achieve the same good results with fewer features than them.

5 Conclusion

In this paper, we propose a new feature selection criterion, called *EVSFSC*, for spectral clustering. *EVSFSC* evaluates the importance of each feature by its impact on the eigenvectors of graph Laplacian with perturbation analysis theory. The extensive experiments demonstrate the excellent performance of our method, compared with four state-of-the-art methods.

Acknowledgements. This work was supported by National Natural Science Foundation of China under Grant No. 60703110.

References

1. Lovasz, L., Plummer, M.: Matching Theory (1986)
2. Wilkinson, J.H.: The Algebraic Eigenvalue Problem Numerical Mathematics and Scientific Computation, Oxford, pp. 62–104 (1988)
3. Joel, N.: Franklin: Matrix Theory (2000)
4. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H. (eds.): Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)
5. Stewart, G.W.: Matrix Algorithms Volumn II: Eigensystems. SIAM, Philadelphia (2001)
6. Hagen, L.W., Kahng, A.B.: New spectral methods for ratio cut partitioning and clustering. IEEE Trans. on CAD of Integrated Circuits and Systems (TCAD) 11(9), 1074–1085 (1992)
7. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell (PAMI) 22(8), 888–905 (2000)
8. Ng, A.Y., Jordan, M.I., Weiss, Y.: On Spectral Clustering: Analysis and an algorithm. In: NIPS 2001, pp. 849–856 (2001)
9. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing (SAC) 17(4), 395–416 (2007)
10. Xiang, T., Gong, S.: Spectral clustering with eigenvector selection. Pattern Recognition (PR) 41(3), 1012–1029 (2008)
11. Huang, L., Yan, D., Jordan, M.I., Taft, N.: Spectral Clustering with Perturbed Data. In: NIPS 2008, pp. 705–712 (2008)
12. Yan, D., Huang, L., Jordan, M.I.: Fast approximate spectral clustering. In: KDD 2009, pp. 907–916 (2009)
13. Hunter, B., Strohmer, T.: Performance Analysis of Spectral Clustering on Compressed, Incomplete and Inaccurate Measurements CoRR abs/1011.0997 (2010)
14. Gong, Y., Huang, T.S.: Incremental Spectral Clustering With Application to Monitoring of Evolving Blog Communities. In: SDM (2007)
15. Ning, H., Xu, W., Chi, Y., Gong, Y., Huang, T.S.: Incremental spectral clustering by efficiently updating the eigen-system. Pattern Recognition (PR) 43(1), 113–127 (2010)
16. Song, X., Zhou, D., Hino, K., Tseng, B.L.: Evolutionary spectral clustering by incorporating temporal smoothness. In: KDD 2007, pp. 153–162 (2007)
17. Coleman, T., Saunderson, J., Wirth, A.: Spectral clustering with inconsistent advice. In: ICML 2008, pp. 152–159 (2008)

18. Wang, X., Davidson, I.: Flexible constrained spectral clustering. In: KDD 2010, pp. 563–572 (2010)
19. Bach, F.R., Jordan, M.I.: Learning Spectral Clustering. In: NIPS (2003)
20. Ozertem, U., Erdogmus, D., Jenssen, R.: Mean shift spectral clustering. Pattern Recognition (PR) 41(6), 1924–1938 (2008)
21. Bhler, T., Hein, M.: Spectral clustering based on the graph p-Laplacian. In: ICML, p. 11 (2009)
22. Kim, Y., Street, W.N., Menczer, F.: Feature selection in unsupervised learning via evolutionary search. In: KDD 2000, pp. 365–369 (2000)
23. Modha, D., Spangler, S.: Feature Weighting in k-Means Clustering. Machine Learning (2002)
24. Huang, J.Z., Ng, M.K., Rong, H., Li, Z.: Automated Variable Weighting in k-Means Type Clustering. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 27(5), 657–668 (2005)
25. Boutsidis, C., Mahoney, M.W., Drineas, P.: Unsupervised Feature Selection for the K-means Clustering Problem. In: NIPS 2009 (2009)
26. Dy, J.G., Brodley, C.E.: Feature Subset Selection and Order Identification for Unsupervised Learning. In: ICML 2000, pp. 247–254 (2000)
27. Law, M.H.C., Jain, A.K., Figueiredo, M.A.T.: Feature Selection in Mixture-Based Clustering. In: NIPS 2002, pp. 625–632 (2002)
28. Roth, V., Lange, T.: Feature Selection in Clustering Problems. In: NIPS 2003 (2003)
29. Jennifer, G.D., Brodley, C.E., Kak, A.C., Broderick, L.S., Aisen, A.M.: Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 25(3), 373–378 (2003)
30. Jennifer, G.D., Brodley, C.E.: Feature Selection for Unsupervised Learning. Journal of Machine Learning Research (JMLR) 5, 845–889 (2004)
31. Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous Feature Selection and Clustering Using Mixture Models. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 26(9), 1154–1166 (2004)
32. Boutemedjet, S., Ziou, D., Bouguila, N.: Unsupervised Feature Selection for Accurate Recommendation of High-Dimensional Image Data. In: NIPS 2007 (2007)
33. Boutsidis, C., Mahoney, M.W., Drineas, P.: Unsupervised feature selection for principal components analysis. In: KDD 2008, pp. 61–69 (2008)
34. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised FeatureSelection Using Feature Similarity. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 24(3), 301–312 (2002)
35. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research (JMLR) 5, 1205–1224 (2004)
36. He, X., Cai, D., Niyogi, P.: Laplacian Score for Feature Selection. In: NIPS 2005 (2005)
37. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: ICML 2007, pp. 1151–1157 (2007)
38. Nie, F., Xiang, S., Jia, Y., Zhang, C., Yan, S.: Trace Ratio Criterion for Feature Selection. In: AAAI 2008, pp. 671–676 (2008)
39. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: KDD 2010, pp. 333–342 (2010)
40. Zhao, Z., Wang, L., Liu, H.: Efficient Spectral Feature Selection with Minimum Redundancy. In: AAAI 2010 (2010)
41. Jiang, Y., Ren, J.: Eigenvalue Sensitive Feature Selection. In: ICML 2011 (2011)