# Feature Selection for Clustering on High Dimensional Data*

Hong Zeng and Yiu-ming Cheung

Department of Computer Science, Hong Kong Baptist University,
Hong Kong SAR, China
{hzeng,ymc}@comp.hkbu.edu.hk

**Abstract.** This paper addresses the problem of feature selection for the high dimensional data clustering. This is a difficult problem because the ground truth class labels that can guide the selection are unavailable in clustering. Besides, the data may have a large number of features and the irrelevant ones can ruin the clustering. In this paper, we propose a novel feature weighting scheme for a kernel based clustering criterion, in which the weight for each feature is a measure of its contribution to the clustering task. Accordingly, we give a well-defined objective function, which can be explicitly solved in an iterative way. Experimental results show the effectiveness of the proposed method.

## 1 Introduction

In many pattern recognition and data mining problems, e.g. the computer vision, text processing and the more recent gene data analysis, etc., the input raw data sets often have a huge number of possible explanatory variables, but there are much fewer samples available. The abundance of variables makes the classification among patterns much harder and less accurate. Under such circumstances, selecting the most discriminative or representative features of a sample inevitably becomes an important issue.

In the literature, most feature selection algorithms have been developed for supervised learning, rather than the unsupervised learning. It is believed that the unsupervised feature selection is more difficult due to the absence of class labels that can guide the search for the relevant information. Until very recently, several algorithms have been proposed to address this issue for clustering. In general, they can be categorized as *wrapper* and *filter* methods according to the evaluation criterion in searching for relevant features. For *wrapper* approaches [5,7], the quality of every candidate feature subset is assessed by investigating the performance of a specific clustering algorithm on this subset, and each candidate subset is obtained by conducting combinatorial search through the space of all feature subsets. These algorithms have shown the success on low dimensional data. Nevertheless, the size of candidate subset space is exponential increased over the number of features. As a result, their computation are laborious, particularly on the high dimensional data. In contrast, the *filter* approaches [6,12,8,4] are more efficient in dealing with the high dimensional data. Such an algorithm first evaluates the

* This work was supported by the Faculty Research Grant of HKBU under Project: FRG/07-08/*II*-54, and the Research Grant Council of Hong Kong SAR under Project: HKBU 210306.

T.-B. Ho and Z.-H. Zhou (Eds.): PRICAI 2008, LNAI 5351, pp. 913–922, 2008.

features by their intrinsic properties (e.g., feature variance, similarity among features, capability of locality preserving, etc.), and then removes a number of less informative features before the clustering. In the literature, the Laplacian score [6] is considered as the state-of-art *filter* method [12]. It selects the features that can preserve the manifold locality described by the weighted nearest neighbor graph, which has a close relationship to the spectral clustering [10,11]. This method has been successfully applied to the real-world high dimensional datasets that possess the manifold characteristics. Nevertheless, it assumes that there should be much less irrelevant features in the data so as to obtain a graph characterizing the authentic similarities among data. In the presence of a large number of irrelevant features, the performance of Laplacian score may be degraded severely.

In this paper, we propose an effective feature selection approach to clustering. The proposed method assigns each feature a real-valued weight to indicate its relevance for the clustering problem, and eventually the issue of feature selection, together with the clustering, is formulated as an optimization problem. Accordingly, we give a kernel based clustering objective function, which can be optimized using an iterative algorithm. In each step of an iteration, the sub-optimization problem is convex and can be easily solved by a well-established optimization technique.

The remainder of the paper is organized as follows. Section 2 introduces the optimization formulation for the feature selection problem, whose solution is given in Section 3. Section 4 presents the extensive experiments on real-world high dimensional datasets. The concluding remarks are given in Section 5.

## 2 The Feature Selection in Clustering as an Optimization Problem

Before giving the feature selection scheme, we first introduce the clustering objective function we proposed in this paper.

### 2.1 The Clustering Objective Function

Let $X = \{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{d\times n}$ denote the data set consisting of $n$ samples over $d$-dimensional space, $S_{ij}(0 \le S_{ij} \le \infty)$ denote the similarity between the points $\mathbf{x}_i$ and $\mathbf{x}_j$, and the similarity matrix $\mathbf{S} = [S_{ij}]_{n\times n}$ is assumed to be symmetric. An intuitive clustering objective is to seek the partition such that the summation of similarities between points in the same cluster is maximized, while that in different clusters is minimized. Such a criterion can be realized to maximize the following cost function:

$$\mathcal{Q}(\mathbb{C}) = \sum_{l=1}^{k} \Big[ \sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathbb{C}_l} (\frac{1}{|\mathbb{C}_l|} - \frac{1}{n}) S_{ij} + \sum_{\mathbf{x}_i,\mathbf{x}_j\notin\mathbb{C}_l} (-\frac{1}{n}) S_{ij} \Big] \tag{1}$$

where $\mathbb{C}$ is a possible partition, $k$ is the number of clusters which is assumed known, $\mathbb{C}_l$ is the set of points contained in the $l$-th cluster ($1 \le l \le k$), and the number of points in the $l$-th cluster is denoted by $|\mathbb{C}_l|$ ($|\mathbb{C}_l| \le n$). The coefficients in front of the similarity items are to defy the effect of different sizes of clusters. Equation (1) can be expressed in a more compact form with matrix operation:

$$\mathcal{Q}(\mathbf{G}) = \sum_{i,j=1}^{n} S_{ij}\tilde{G}_{ij} = trace(\mathbf{S}\tilde{\mathbf{G}}) \tag{2}$$

where $\tilde{\mathbf{G}} = \mathbf{\Pi}_n\mathbf{G}\mathbf{\Pi}_n$, $\mathbf{\Pi}_n \in \mathbb{R}^{n\times n}$ is the centering matrix defined as $\mathbf{\Pi}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{e}_n\mathbf{e}_n^T$, and $\mathbf{e}_n$ is a vector of all ones of size $n$. $\mathbf{G} \in \mathbb{R}^{n\times n}$ is a matrix whose entry is defined as: $G_{ij} = \frac{1}{|\mathbb{C}_l|}$, if $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{C}_l$, and zero otherwise. If we define the hard cluster indicator matrix $\mathbf{L} \in \mathbb{R}^{n\times k}$ as: $L_{il} = |\mathbb{C}_l|^{-\frac{1}{2}}$,if $\mathbf{x}_i \in \mathbb{C}_l$ and zero otherwise. It is easy to verify that the cluster indicator matrix satisfies: $\mathbf{L}\mathbf{L}^T = \mathbf{G},\ \mathbf{L}^T\mathbf{L} = \mathbf{I}_k$, where $\mathbf{I}_k \in \mathbb{R}^{k\times k}$ is the unit matrix. Then Equation (2) can be expressed as:

$$\mathcal{Q}(\mathbf{L}) = trace(\mathbf{S}\mathbf{\Pi}_n\mathbf{L}\mathbf{L}^T\mathbf{\Pi}_n) = trace(\mathbf{L}^T\mathbf{\Pi}_n\mathbf{S}\mathbf{\Pi}_n\mathbf{L}). \tag{3}$$

By the spectral relaxation [1], we allow $L_{ij}$ to take a continuous value, subject to the constraint $\mathbf{L}^T\mathbf{L} = \mathbf{I}_k$ so as to turn it into a tractable continuous optimization problem. Hence, the clustering criterion can be formulated as:

$$\max_{\mathbf{L}^T\mathbf{L}=\mathbf{I}_k} trace(\mathbf{L}^T\mathbf{\Pi}_n\mathbf{S}\mathbf{\Pi}_n\mathbf{L}). \tag{4}$$

### 2.2 The Weighting Scheme to Select Features

The matrix $\mathbf{S}$ is not necessarily fixed. In fact, we select the relevant features to enhance the similarity matrix for maximizing the criterion in Equation (4), i.e.

$$\max_{\mathbf{S},\mathbf{L}}\ \mathcal{Q}(\mathbf{L},\mathbf{S}) = trace(\mathbf{L}^T\mathbf{\Pi}_n\mathbf{S}\mathbf{\Pi}_n\mathbf{L})\ \ s.t.\ \mathbf{L}^T\mathbf{L} = \mathbf{I}_k. \tag{5}$$

Suppose the RBF kernel function is adopted as the similarity, i.e., $S_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\sum_{p=1}^{d}(\mathbf{x}_i^{(p)} - \mathbf{x}_j^{(p)})^2}{t})$, where $\mathbf{x}_i^{(p)}$ is the value of the $p$-th feature for the point $\mathbf{x}_i$. Let $w_p \in \{0,1\}$ $(1 \le p \le d)$ be the relevance indicator value associated with the $p$-th feature, i.e., $w_p = 1$ if the $p$-th feature is selected to form the relevant feature subset and 0 otherwise. Therefore, a natural feature selection scheme is to use a modified similarity matrix, denoted as $\mathbf{S}^w$, whose $(i,j)$-th element is defined as:

$$S_{ij}^w = K(\mathbf{w}\circ\mathbf{x}_i, \mathbf{w}\circ\mathbf{x}_j) = e^{-\frac{\sum_{p=1}^{d}\mathbf{w}_p(\mathbf{x}_i^{(p)} - \mathbf{x}_j^{(p)})^2}{t}}, \tag{6}$$

where $\circ$ denotes the element-wise multiplication. Since maximization of $\mathcal{Q}(\mathbf{L}, \mathbf{S}^w)$ for all possible feature subsets is infeasible for high dimensional data, we relax the indicator $w_p$ to real-valued nonnegative weight (i.e. $w_p \ge 0$), and a large value of $w_p$ will indicate that the $p$-th feature is more important to the similarity formation. It is observed that substituting the modified similarity matrix $\mathbf{S}^w$ into (5) will lead to a nonlinear optimization problem with respect to $\mathbf{w}$, which is very difficult to solve, due to the nonlinearity introduced by the RBF kernel function. In order to overcome this difficulty, we propose a simple but effective weighting scheme:

$$S_{ij}^w = \sum_{p=1}^{d} w_p K(\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)}) = \sum_{p=1}^{d} w_p e^{-\frac{(\mathbf{x}_i^{(p)} - \mathbf{x}_j^{(p)})^2}{t}} = \sum_{p=1}^{d} w_p K_p(\mathbf{x}_i, \mathbf{x}_j), \tag{7}$$

where we define $K_p(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_i^{(p)}, \mathbf{x}_j^{(p)})$ as the $(i,j)$-th element of the kernel matrix $\mathbf{K}_p$ that is constructed by using only the $p$-th feature of the data points. Furthermore, we normalize $\mathbf{K}_p$ to $\mathbf{K}_p \leftarrow \mathbf{D}_p^{-\frac{1}{2}}\mathbf{K}_p\mathbf{D}_p^{-\frac{1}{2}}$, where $\mathbf{D}_p$ is a diagonal matrix with the row sum of $\mathbf{K}_p$ in the diagonal, and the operation "$\mathbf{A} \leftarrow \mathbf{B}$" means that the value of $\mathbf{B}$ is assigned to $\mathbf{A}$. The normalized similarity can be interpreted as the probability of $\mathbf{x}_i^{(p)}$ (or $\mathbf{x}_j^{(p)}$) being close to $\mathbf{x}_j^{(p)}$ (or $\mathbf{x}_i^{(p)}$). For fixed $\mathbf{w}$, the aggregated and normalized $\mathbf{K}_p$ form a combination $\mathbf{S}^w = \sum_{m=1}^{d} w_p\mathbf{K}_p$. If $w_p = 0$, this implies that $\mathbf{S}^w$ will not depend on the original $p$-th feature. Obviously, $\mathbf{S}^w$ is still symmetric and has nonnegative elements. Subsequently, maximizing the criterion in (5) finally becomes the following optimization problem:

$$\max_{\mathbf{w},\mathbf{L}} \mathcal{Q}(\mathbf{L}, \mathbf{w}) = \max_{\mathbf{w},\mathbf{L}} \; trace\big[\mathbf{L}^T\mathbf{\Pi}_n(\sum_p w_p\mathbf{K}_p)\mathbf{\Pi}_n\mathbf{L}\big]$$
$$s.t.\; w_p \geq 0, \|\mathbf{w}\|^2 = 1, \qquad \mathbf{L}^T\mathbf{L} = \mathbf{I}_k, \tag{8}$$

where the constraint $\|\mathbf{w}\|^2 = 1$ prevents the maximization from increasing without bound.

## 3 The Solution to the Optimization Problem

The objective function in Equation (8) is not convex. However, if one of the two components ($\mathbf{w}$ and $\mathbf{L}$) is fixed, the objective function will be convex in terms of the other component. Subsequently, the optimization problem becomes easy to solve, which will enable us to solve the problem by updating $\mathbf{w}$ and $\mathbf{L}$ iteratively to find a (local) optimal solution for (8).

### 3.1 Calculation of L for a Given w

Given a weight vector $\mathbf{w}$, the maximization problem specified in Equation (8) reduces to the following trace maximization problem:

$$\max_{\mathbf{L}} \; trace\big(\mathbf{L}^T\tilde{\mathbf{K}}\mathbf{L}\big) \;\; s.t.\; \mathbf{L}^T\mathbf{L} = \mathbf{I}_k. \tag{9}$$

where $\tilde{\mathbf{K}}$ is defined as $\tilde{\mathbf{K}} = \mathbf{\Pi}_n(\sum_p w_p\mathbf{K}_p)\mathbf{\Pi}_n$. According to the Ky Fan theorem [2], an optimal solution for $\mathbf{L}$ is given by the $k$ eigenvectors of $\tilde{\mathbf{K}}$ corresponding to the $k$ largest eigenvalues, where $k$ is the number of clusters.

### 3.2 Calculation of w as Given L

Given a cluster indicator matrix $\mathbf{L}$, the maximization problem specified in Equation (8) reduces to:

$$\max_{\mathbf{w}} \; \sum_p w_p trace(\mathbf{L}^T\mathbf{\Pi}_n\mathbf{K}_p\mathbf{\Pi}_n\mathbf{L}) \;\; s.t.\; w_p \geq 0, \|\mathbf{w}\|^2 = 1. \tag{10}$$

Let $z_p = trace(\mathbf{L}^T \mathbf{\Pi}_n \mathbf{K}_p \mathbf{\Pi}_n \mathbf{L})$. Intuitively, it can be interpreted as the contribution of the $p$-th feature to the clustering. Then (10) can be simplified as:

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{z}, \;\; s.t.\; w_p \geq 0, \|\mathbf{w}\|^2 = 1. \tag{11}$$

Applying the Lagrangian method to the optimization problem (11), we can obtain its analytical solution $\mathbf{w} = (\mathbf{z})^+ / \|(\mathbf{z})^+\|_2$, where $(z_p)^+ = \max(z_p, 0)$. Fortunately, $z_p$ is always nonnegative because $z_p = trace(\mathbf{L}^T \mathbf{\Pi}_n \mathbf{K}_p \mathbf{\Pi}_n \mathbf{L}) = \sum_{l=1}^{k} (\mathbf{\Pi}_n \mathbf{L}_l)^T \mathbf{K}_p (\mathbf{\Pi}_n \mathbf{L}_l)$ $\geq 0$, which follows from the positive semi-definite property of $\mathbf{K}_p$. Therefore, given $\mathbf{L}$, we can obtain a global maximizer for (10), i.e., $\mathbf{w} = \mathbf{z}/\|\mathbf{z}\|_2$. Namely, the weight for each feature is proportional to its contribution to the overall clustering quality.

### 3.3 The Main Algorithm

Based on the discussion described above, we propose to develop an iterative algorithm. The pseudo-code of the main algorithm is given in Algorithm 1. The final discrete

**input** : $\mathbf{X}$ , $k$, $\epsilon$
**output**: $\mathbf{L}$, $\mathbf{w}$

1 Construct $\mathbf{K}_p (p = 1, \ldots, d)$ with RBF kernel function using only the $p$-th row of $\mathbf{X}$, and normalize each kernel matrix as:$\mathbf{K}_p \leftarrow \mathbf{D}_p^{-\frac{1}{2}} \mathbf{K}_p \mathbf{D}_p^{-\frac{1}{2}}$, where $\mathbf{D}_p$ is a diagonal matrix with the row sum of $\mathbf{K}_p$ in the diagonal;
2 Set the initial weight vector $\mathbf{w}$ to $\mathbf{e}_d / \|\mathbf{e}_d\|_2$;
3 **while** *the **relative** change of the objective function value* $\geq \epsilon$ **do**
4 Update $\mathbf{L}$ as in Section 3.1;
5 Update $\mathbf{w}$ as in Section 3.2;
6 Record the objective function value in (8);
7 **end**
8 return $\mathbf{L}$, $\mathbf{w}$;

**Algorithm 1.** The main algorithm for the integrated feature selection in clustering

clustering result can be obtained by applying $k$-means on the rows of the relaxed cluster indicator matrix $\mathbf{L}$ as in [10]. The convergence of the proposed algorithm is guaranteed, as shown in Theorem 1.

**Theorem 1.** *The proposed algorithm always converges.*

*Proof.* Given an arbitrary weight vector $\mathbf{w}^*$ that satisfies the required constraints, we can obtain that:

$$\begin{aligned}
&\max_{\mathbf{L}} \mathcal{Q}(\mathbf{L}, \mathbf{w}^*) = \max_{\mathbf{L}} trace\big[\mathbf{L}^T \mathbf{\Pi}_n (\sum_p w_p^* \mathbf{K}_p) \mathbf{\Pi}_n \mathbf{L}\big] = \lambda_1 + \cdots + \lambda_k \\
&\leq \lambda_1 + \cdots + \lambda_n = trace\big[\mathbf{\Pi}_n (\sum_p w_p^* \mathbf{K}_p) \mathbf{\Pi}_n\big] \\
&= \sum_p w_p^* trace(\mathbf{\Pi}_n \mathbf{K}_p \mathbf{\Pi}_n) \leq \max_{w_p^* \geq 0, \|\mathbf{w}^*\|^2 = 1} \mathbf{w}^{*T} \nu = \|\nu\|_2,
\end{aligned} \tag{12}$$

where $\nu_p = trace(\mathbf{\Pi}_n \mathbf{K}_p \mathbf{\Pi}_n)$ is fixed. According to (12), the maximum of (8) is always upper bounded by a fixed finite value. Since Step 4 and 5 of the main algorithm optimize the same objective function in (8), its value is non-decreasing, the algorithm will always converge.

In the implementation, we set the threshold $\epsilon$ at $0.0005$ for checking the convergence. We observe from our experiments that the proposed algorithm converges in less than 10 iterations, and typically within 3 to 4 iterations.

## 4 Experiments

We conducted several experiments to demonstrate the effectiveness of the proposed algorithm. Five benchmark data sets were used in our experiments, and their characteristics are summarized in Table 1. On each data set, we investigated whether our integrated feature selection and clustering algorithm (denoted as *integrated*) could improve the conventional spectral clustering algorithm, normalized cut (Ncut)[1] [11], which is equivalent to optimize the same clustering criterion in (4) (given the row-sum normalized $\mathbf{S}$ matrix is fixed), but does not consider the feature selection issue (denoted as *nofs+Ncut*). Furthermore, we compared our feature selection method with the state-of-art unsupervised one, the Laplacian score [6], whose starting point is to seek the feature that can preserve the locality. Specifically, it essentially ranks the features according to the following criterion in the ascending order [6]:$\sum_{ij} w_{ij}(\mathbf{x}_i^{(p)} - \mathbf{x}_j^{(p)})^2$, where $w_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ is the local similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, obtained with all the features. Since the Laplacian score is a ranking-based *filter* approach, it does not perform the clustering. For a fair comparison, the clustering result obtained by our integrated method was not utilized when comparing with the Laplacian score. Instead, we ranked the features according to their weights, the performance of the normalized cut clustering with the top-ranked features was then used to evaluate our weighting scheme (denoted as *wrank+Ncut*) and the Laplacian score (denoted as *laprank+Ncut*).

**Table 1.** Summary of the benchmark data sets

| Data Set | #DIM ($p$) | #INST ($n$) | #CL ($k$) |
|---|---|---|---|
| PIEC27_5p | 1280 | 105 | 5 |
| LEUKEMIA | 999 | 38 | 2 |
| LUNG | 1000 | 197 | 4 |
| MULTITISSUE | 1000 | 103 | 4 |
| ST.LEUKEMIA | 985 | 248 | 6 |

As we have the label information of all five benchmark datasets, the clustering results were evaluated by comparing the obtained label of each data points with the ground truth. We used two standard measurements: the accuracy (ACC) and the normalized mutual information (NMI), higher values for both measurements will indicate good clustering performance.

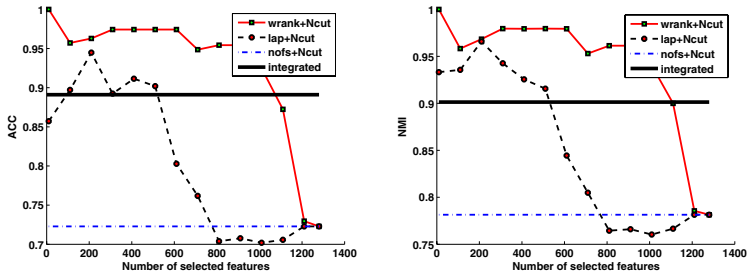
[1] The source code in MATLAB is downloaded from: http://www.cis.upenn.edu/∼jshi/software/

**Fig. 1.** Sample images from PIE face database under varying illumination conditions

Throughout the experiments, the parameter $t$ in RBF kernel function was simply set at $0.0025 * \max(B^{(p)})$ for our method, where $\max(B^{(p)})$ is the maximum squared pairwise Euclidean distance between the elements in the $p$-th feature. The similarity used by the Ncut algorithm was also built with RBF kernel function, and the parameter $t$ in RBF kernel function was set at $0.0025 * \max(B)$, where $\max(B)$ is the maximum squared pairwise Euclidean distance between the points.

### 4.1 Faces with Varying Illumination Conditions

A subset of the CMU PIE face database was used in this experiment. It contains 68 human subjects of the frontal poses (C27) but under different illumination conditions, with each subject having 21 faces. We used the cropped images[2] of $32 \times 32$ pixels. Samples extracted from the database are represented in Figure 1, in which it is observed that partitioning the face images of the same person in an unsupervised manner may be difficult because different persons appear similar under varying illumination conditions from the viewpoint of the image intensity. It therefore suggests the illumination insensitive features, e.g., the image gradient, may help the discrimination among identities [3]. In this experiment, we simply used the wavelet transform, which is able to compute the gradient, to generate the features for an image. However, a large amount of the wavelet coefficients are irrelevant for the task of separating between facial identities and it is therefore the goal of our algorithm to find those relevant coefficients as well as a more accurate clustering.




(a) (b)

**Fig. 2.** Clustering results on PIE C27 data. (a): Comparison using the ACC index. (b): Comparison using the NMI index.

[2] The data is obtained from http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html

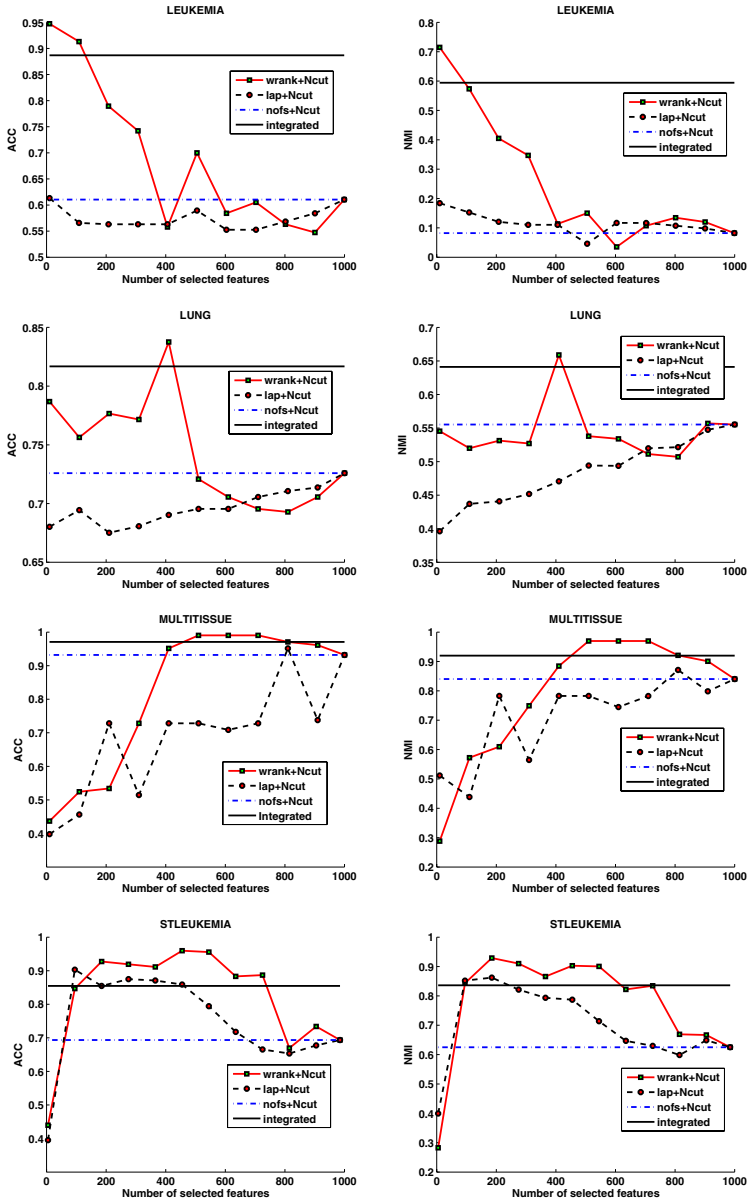



**Fig. 3.** Clustering results on the gene expression datasets. The first figure in each row demonstrates the comparison using the ACC index, the second one denotes the comparison using the NMI index.

Specifically, we first scaled the pixel value into $[0, 1]$, then the level-1 and level-2 Haar wavelet decompositions were performed over an image. After respectively normalizing the wavelet coefficients from each level by the average value in each corresponding level, we cascaded them to form a 1280-dimension vector representation for an image. Data from 5 classes were randomly selected out of the 68 objects, and this process was repeated 20 times, and the average performance was reported for all algorithms on the same data. Experimental results are summarized in Figure 2. It is observed that the proposed integrated algorithm significantly improves the conventional Ncut clustering with no feature selection. Besides, our weighting scheme largely outperforms the Laplacian score in the presence of many irrelevant wavelet coefficients.

### 4.2 Gene Expression Data

In this experiment, we studied the feature selection for the clustering on four public gene expression datasets: LEUKEMIA, LUNG, MULTITISSUE, STLEUKEMIA[3]. The characteristics of these datasets are summarized in Table 1. For LEUKEMIA dataset, expression values were first thresholded with a floor of 1 and a ceiling of 16000, followed by a base 10 logarithmic transformation. Then each gene was standardized to zero mean and unit variance across samples. For the MULTITISSUE and STLEUKEMIA datasets, each gene was standardized. For LUNG dataset, the already preprocessed expression profiles was used.

The average performance of all the algorithms was reported over 10 runs. The results are summarized in Figure 3. It can be seen that the integrated feature selection and clustering algorithm consistently outperforms the Ncut clustering with all features on all the datasets. In Figure 3, the Laplacian score does not always improve the clustering with no feature selection, and even falls much behind (see results for LUNG, MULTITISSUE), i.e., it is not robust for the gene expression data with large amount of noisy and irrelevant features. On the contrary, our weighting method demonstrates robustness against such features. A possible explanation for this superiority is that, for the gene expression data containing many irrelevant features, it may not be sensible to select features that try to preserve the locality, while selecting features that directly help the class discovery may be more appropriate.

## 5 Concluding Remarks

In this paper, we have proposed a novel feature weighting scheme for kernel-based clustering. It has a clearly defined objective function and can be solved iteratively. Rather than relying on the possible spurious "intrinsic" properties that may have been corrupted by irrelevant features, the weight assigned to each feature has direct relation to the clustering task. Experimental results have shown that our integrated feature weighting and clustering algorithm consistently improves the conventional kernel-based clustering without considering the feature selection. Moreover, when the feature weighting is used as a feature ranking method, it outperforms the state-of-art unsupervised one, the Laplacian score, in the presence of a lot of irrelevant features.

[3] The gene expression datasets are obtained from [9].

It is necessary to point out that the performance of the integrated algorithm (*integrated*) is generally inferior to the best performance of normalized cut with the most relevant features selected by our weighting scheme (*wrank+Ncut*). The reason is that, the clustering quality of the integrated method essentially depends on the similarity matrix formed by a linear combination of individual kernel matrix constructed from each feature, but the interaction among features has not been presented. In contrast, in *wrank+Ncut*, the enhanced similarity matrix formed with the most relevant features selected by our weighting, does not omit the correlation among them. However, the most difficult issue for the ranking based approach in clustering is the determination of the number of relevant features to be selected. Since the cross-validation, a commonly used model selection technique in supervised learning, cannot be directly applied in unsupervised environment, in which the ground truth class labels are unavailable. Therefore the *integrated* method may be superior to the *wrank+Ncut* from the practical viewpoint. It will be the future work to incorporate the nonlinear interaction among features into our algorithm.

## References

1. Zha, H., He, X., Ding, C., Simon, H.: Spectral relaxation for k-means clustering. In: NIPS, pp. 1057–1064 (2001)
2. Fan, K.: On a theorem of Weyl concerning eigenvalues of linear transformations. In: PNAS, pp. 652–655 (1949)
3. Chen, H.F., Belhumeur, P.N., Jacobs, D.W.: In search of illumination invariants. In: CVPR, pp. 254–261 (2000)
4. Dash, M., Choi, K., Scheuermann, P., Liu, H.: Feature selection for clustering-a filter solution. In: ICDM, pp. 115–122 (2002)
5. Dy, J., Brodley, C.: Feature selection for unsupervised learning. JMLR 5, 845–889 (2004)
6. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS, pp. 507–514 (2005)
7. Law, M.H., Jain, A.K., Figueiredo, M.A.T.: Feature selection in mixture-based clustering. In: NIPS, pp. 609–616 (2002)
8. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature felection using feature similarity. PAMI 24(3), 301–312 (2002)
9. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52(1), 91–118 (2003)
10. Ng, A., Jordan, M., Weiss, Y.: On spectral slustering: analysis and an algorithm. In: NIPS (2002)
11. Yu, S.X., Shi, J.: Multiclass spectral clustering. In: ICCV, pp. 313–319 (2003)
12. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: ICML, pp. 1151–1158 (2007)