Regularized and Constrained Self-representation for Robust Feature Selection

Anonymous Author(s)

Abstract

Feature selection is an important topic in data mining. In this paper, we focus on the problem in unsupervised scenario, which is challenging due to the absence of labels. We formulate our model RRCS from the viewpoint of self-representation. For the selection matrix, unlike many previous methods which take the $\ell_{2,1}$-norm regularization to avoid trivial solution and achieve feature selection, we directly use the $\ell_{2,0}$-norm constraint to obtain a more accurate solution. By explicitly considering the representation residue, we relax the hard linear constraint in self-representation, making our model better deal with the nonlinear case. Using the $\ell_{2,1}$-norm loss term, the robustness of RRCS is achieved. Moreover, we add a graph regularization to preserve the local structure of the original data. An efficient algorithm is derived to solve the regularized and constrained problem. Extensive experiments on several datasets demonstrate the effectiveness of the proposed method.

Introduction

In many areas such as data mining, machine learning, and biological study, one is often confronted with high dimensional data. Dealing with these data directly is both time and memory consuming, and may degenerate the performance of learning algorithms, since there usually exist irrelevant and redundant features and noise in the original data. As a data preprocessing technique, dimensionality reduction can be mainly categorized into feature extraction and feature selection. PCA (Turk and Pentland 1991), LDA (Belhumeur *et al.* 1997) and LLE (Roweis and Saul 2000), just to name a few, are classical feature extraction methods that project the data into a new space with lower dimensionality, without informing us on which features are important. Feature selection aims to select a subset from the original features for a more compact representation. In some cases where the features have natural meanings, such as selecting a few of genes associated with a given disease or biological function in DNA microarray analysis, and selecting some key words in text mining, feature selection is preferred. Collecting the labeled data is usually expensive, while the unlabeled data is abundant and can be easily obtained. Therefore, it is of great value to study the unsupervised feature selection problem, which is the focus of this paper.

Sparsity regularization technique, due to its solid theoretical foundation and superior properties, has received growing attention in the studies of feature selection in recent years. For a learning model, the norm of regression coefficients is usually added to the objective function as a regularization term to avoid over-fitting. It can be viewed that the model is actually doing feature selection when the norm leads some coefficients to be zero. The common idea of the regularization based feature selection methods is to choose a proper norm for the selection matrix (vector) to achieve sparseness. The features corresponding to the non-zero rows (entries) are then removed. In general, the ℓ_0-norm, ℓ_1-norm and ℓ_∞-norm are used to achieve flat sparsity, while the $\ell_{2,0}$-norm, $\ell_{2,1}$-norm and $\ell_{2,\infty}$-norm are for structural sparsity.

Unsupervised feature selection is challenging due to the absence of labels, and there are three main categories of the regularization based methods. A natural way is to transform the unsupervised problem into a fake supervised one by producing the pseudo labels with spectral analysis (Cai *et al.* 2010; Zhao *et al.* 2010). The works in (Hou *et al.* 2011; Li *et al.* 2012; Qian and Zhai 2013; Shi *et al.* 2014; Qian and Zhai 2015) share the similar idea but embed the pseudo label generation and feature selection into a joint framework. Another direction is based on structure or similarity preservation that a graph is usually exploited to characterize the discriminative or geometrical information of the data (Masaeli *et al.* 2010a; Gu *et al.* 2011; Yang *et al.* 2011; Zhao *et al.* 2013; Du *et al.* 2013; Liu *et al.* 2014; Tang *et al.* 2014; Nie *et al.* 2016). The third one is based on self-representation, with the idea that the selected features should be able to reconstruct each feature by linear combination. (Masaeli *et al.* 2010b; Zhu *et al.* 2015; Zhao *et al.* 2015).

For multi-class feature selection problems, the $\ell_{2,0}$-norm constraint is the most desired one to achieve structural sparsity. However, in the methods mentioned above, the most popular way to do feature selection is by the $\ell_{2,1}$-norm regularization. Since the $\ell_{2,0}$-norm is non-convex and hard to tackle, and the convex $\ell_{2,1}$-norm regularization is considered to be approximately identical to the $\ell_{2,0}$-norm regularization. (Cai *et al.* 2013) solved the problem with a $\ell_{2,0}$-norm constraint, and (Zhang *et al.* 2014) adopted a general $\ell_{2,p}$-

norm ($0 \leq p \leq 1$) regularization, while these two are for supervised case. Another problem for many existing methods is that in the real data, noise is always there and it may have adverse effect on the constructed graph. Self-representation based methods naturally consider feature redundancy, while the underlying assumption of linear correlation between features does not always hold in real applications. To address these problems, in this paper, we are going to propose a Robust, Regularized and Constrained Self-representation (RRCS) model for unsupervised feature selection. The contributions of the paper are mainly four-fold:

- We use the $\ell_{2,0}$-norm constraint, rather than its regularization form or other approximated regularization terms for unsupervised feature selection. The number of selected features can be set directly instead of by tuning the parameter.
- By explicitly considering the representation residue, we relax the hard linear constraint in self-representation, making our model better deal with the case that the correlation between features is non-linear.
- To deal with the noise and outliers, we take the robust $\ell_{2,1}$-norm loss term. The residue can also be explained as a modeling of noise and outliers, and the regularization of the residue tends to suppress their effect.
- A simple algorithm is derived to solve the robust, regularized and constrained model. Extensive experiments on several datasets show that our approach is effective.

Background and Related Works

Notations

Let $X \in \mathbb{R}^{n \times d}$ be the data matrix, where n is the number of samples and d is the dimensionality. The trace of matrix is denoted by $Tr(\cdot)$. I_n is an identity matrix of size $n \times n$. For matrix $W = \{W_{ij}\}$, the i-th row and the j-th column are denoted by bold lowercase $\mathbf{w}_i$ and $\mathbf{w}^j$, respectively. The ℓ_p-norm of vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_p = (\sum_{i=1}^{n} |v_i|^p)^{\frac{1}{p}}$ ($p \neq 0$). Thus the ℓ_∞-norm of $\mathbf{v}$ is $\|\mathbf{v}\|_\infty = \max_i |v_i|$. The $\ell_{r,p}$-norm of $W \in \mathbb{R}^{d \times m}$ is defined as $\|W\|_{r,p} = (\sum_{i=1}^{d}(\sum_{j=1}^{m} |W_{ij}|^r)^{\frac{p}{r}})^{\frac{1}{p}}$ ($r \neq 0, p \neq 0$). The $\ell_{2,0}$-norm of W is the number of non-zero rows of W. The $\ell_{2,2}$-norm of W is exactly its Frobenius norm. Note that the $\ell_{2,0}$-norm is not a valid norm since it does not satisfy the positive scalability: $\|\lambda W\|_{2,0} = |\lambda| \|W\|_{2,0}$ for scalar λ. The term "norm" here is just for convenience.

Self-representation

self-representation is feature-level representation while sparse representation and low-rank representation belong to sample-level representation (Zhu *et al.* 2015). Without labels, self-representation simply uses the features as response, under the assumption that each feature can be linearly represented by all features. For the i-th feature $\mathbf{x}^i \in \mathbb{R}^n$:

$$\mathbf{x}^i \approx \sum_{j=1}^{d} \mathbf{x}^j W_{ji} = X\mathbf{w}^i. \tag{1}$$

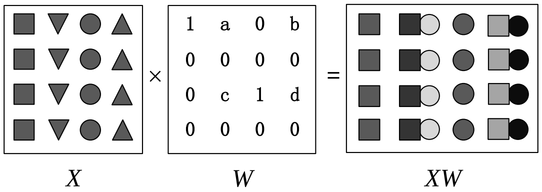


Figure 1: An illustration of self-representation based feature selection. Different shapes in X mean different features, *i.e.*, each column is a feature. a, b, c and d are representation coefficients. The first and the third feature are selected, then the second and the fourth feature in XW are represented by the linear combination of the selected features. Here we distinguish different combinations by grayscale.

where $\mathbf{w}^i = [W_{1i}, \ldots, W_{di}]^T$ is the representation coefficient vector of $\mathbf{x}^i$. Then for all features:

$$X \approx XW. \tag{2}$$

W is the representation coefficient matrix, or it can be called the selection matrix when doing feature selection. We can see that if the i-th feature is selected, then W_{ii} should be equal to one and the other entries of $\mathbf{w}^i$ are all zeros. And if the entries in the j-th row of W are all zeros, then the j-th feature is not selected and it will be reconstructed by the linear combination of the selected features, which accounts for feature redundancy. The number of zero rows of W corresponds to the number of selected features. We conceptually illustrate above analysis in Figure 1.

Related Works

CPFS CPFS (Masaeli *et al.* 2010b) is developed from the viewpoint of PCA (Turk and Pentland 1991), whose goal is to maximize the variance, or equivalently, to minimize the reconstruction error of self-representation. The objective function of CPFS is:

$$\min_{W} \|X - XW\|_F^2 + \lambda \sum_{i=1}^{d} \|\mathbf{w}_i\|_\infty, \tag{3}$$

CPFS achieves the flat sparsity by forcing the biggest entry in each row of W to be small. parameter λ controls the tradeoff between representation residue and sparsity. Ranging λ from zero to infinity means ranging the number of selected features from d to 0. Therefore, if we want to select k features, we need to tune λ to make the number of non-zero rows of W exactly equal to k.

RSR Considering that there usually exist some outliers in the data matrix, and the Frobenius norm is sensitive to outliers, RSR (Zhu *et al.* 2015) aims to solve the following joint $\ell_{2,1}$-norm problem:

$$\min_{W} \|X - XW\|_{2,1} + \lambda \|W\|_{2,1}. \tag{4}$$

The regularization term is to avoid trivial solution and achieve feature selection. Similar to CPFS, large value of λ means selecting fewer features. Actually, RSR can be viewed as an unsupervised version of RFS (Nie *et al.* 2010a). The limitation of RSR is that in some cases, the features are independent, or the correlation between features is nonlinear, then self-representation perhaps does not work well.

GRFS GRFS (Zhao *et al.* 2015) considers the representation coefficient matrix A and the feature selection matrix Λ separately. ΛA plays the role of W in CPFS and RSR. The objective function of GRFS is as follows:

$$\begin{aligned} &\min_{A,\Lambda} \|X - X\Lambda A\|_F^2 + \beta Tr(\Lambda X^T L X \Lambda), \\ &s.t. \quad \Lambda = diag(\boldsymbol{\lambda}), Card(\boldsymbol{\lambda}) = k, \lambda_i \in \{0, 1\} \end{aligned} \tag{5}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a binary vector and its cardinality is k. Λ is a diagonal matrix with the i-th diagonal entry being λ_i. The second term is the graph regularization which preserves the intrinsic structure in the original data space. This problem is computationally intractable due to the integer variables in vector $\boldsymbol{\lambda}$, thus the authors tried to optimize the relaxed one:

$$\min_{A,\Lambda} \|X - X\Lambda A\|_F^2 + \beta Tr(\Lambda X^T L X \Lambda) + \alpha \|\boldsymbol{\lambda}\|_1. \tag{6}$$

However, this relaxed problem exists trivial solution.

The Proposed Method

Formulation

We rewrite the self-representation model in Eq.(2) as follows:

$$X = XW + Z, \tag{7}$$

where $Z \in \mathbb{R}^{n \times d}$ is the representation residue. The previous methods usually use a matrix norm to characterize the residue such as the term $\|Z\|_{2,1} = \|X - XW\|_{2,1}$ in RSR. The hard linear constraint in self-representation may be overstrict for the features that are not linearly correlated. Motivated by the recent work Flexible Manifold Embedding (FME) (Nie *et al.* 2010b) for feature extraction, we aims to consider the residue Z explicitly by solving the following problem:

$$\min_{Z, \|W\|_{2,0} = k} \|X - XW - Z\|_{2,1} + \beta R(Z). \tag{8}$$

We relax the linear constraint by introducing the residue into the objective function to better cope with the nonlinear case. The selection matrix W and the representation residue Z are optimized simultaneously. $R(Z)$ is a kind of regularization on Z to avoid trivial solution. To obtain a more accurate solution for feature selection, we directly use the $\ell_{2,0}$-norm of W as a constraint. Parameter k has explicit meaning. Thus the number of selected features can be set directly instead of by tuning the parameter.

Here Z models the mismatch between X and XW. From another viewpoint, the residue Z may be considered as a modeling of the noise and outliers, and the regularization term $R(Z)$ tends to suppress their effect. Following FME, we simply use the Frobenius norm of Z as $R(Z)$. Further, we would like the reconstructed data to preserve the local structure of the original data. A common way (Cai *et al.* 2007; Nie *et al.* 2010b; Zhao *et al.* 2015) is by minimizing:

$$\frac{1}{2} \sum_{i,j=1}^{n} \|\mathbf{x}_i W - \mathbf{x}_j W\|_2^2 S_{ij} = Tr(W^T X^T L X W), \tag{9}$$

where $S \in \mathbb{R}^{n \times n}$ is the symmetric similarity matrix that S_{ij} represents the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. The graph Laplacian $L \in \mathbb{R}^{n \times n}$ is defined as $L = D - S$ and the normalized one is $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, where D is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$. Therefore, the final objective function is:

$$\begin{aligned} &\min_{W,Z} \|X - XW - Z\|_{2,1} + \alpha Tr(W^T X^T L X W) \\ &+ \beta \|Z\|_F^2 \quad s.t. \quad \|W\|_{2,0} = k. \end{aligned} \tag{10}$$

Algorithm 1 ALM Method to Solve Eq.(11)

```
Set $1 < \rho < 2$, initialize $\mu > 0$,$\Lambda$.
while not converge do
    Update $X$ by $\min_X f(X) + \frac{\mu}{2}\|h(X) + \frac{1}{\mu}\Lambda\|_F^2$.
    Update $\Lambda$ by $\Lambda := \Lambda + \mu h(X)$.
    Update $\mu$ by $\mu := \rho\mu$.
end while
```

Optimization

In this subsection, we are going to optimize the robust, regularized and constraint problem based on the Augmented Lagrange Multiplier (ALM) method (Bertsekas 1982).

A Brief Description of ALM For the following constrained problem:

$$\min_X f(X) \quad s.t. \quad h(X) = 0, \tag{11}$$

we can transform it into an unconstrained one based on ALM:

$$\begin{aligned} &\min_{X,\Lambda,\mu} f(X) + Tr(\Lambda^T h(X)) + \frac{\mu}{2}\|h(X)\|_F^2 \\ \Leftrightarrow &\min_{X,\Lambda,\mu} f(X) + \frac{\mu}{2}\|h(X) + \frac{1}{\mu}\Lambda\|_F^2, \end{aligned} \tag{12}$$

where Λ is the Lagrange multiplier, and μ is the quadratic penalty parameter. The procedure of ALM to solve Eq.(11) is described in Algorithm 1.

Solving Eq.(10) Using ALM Method We introduce two slack variables $E \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{d \times d}$, Eq.(10) can then be rewritten as:

$$\begin{aligned} &\min_{W,Z,E,V} \|E\|_{2,1} + \alpha Tr(W^T X^T L X W) + \beta \|Z\|_F^2 \\ &s.t. \quad E = X - XW - Z, W = V, \|V\|_{2,0} = k \end{aligned} \tag{13}$$

Based on ALM, we need to solve:

$$\begin{aligned} &\min_{W,Z,E,V} \|E\|_{2,1} + \frac{\mu}{2}\|E - X + XW + Z + \frac{1}{\mu}\Lambda\|_F^2 \\ &+ \frac{\mu}{2}\|W - V + \frac{1}{\mu}\Sigma\|_F^2 + \alpha Tr(W^T X^T L X W) \\ &+ \beta \|Z\|_F^2 \quad s.t. \quad \|V\|_{2,0} = k \end{aligned} \tag{14}$$

A joint minimization with respect to the four variables is difficult. Thus we divide the problem into four subproblems, and optimize them alternatively and iteratively. The whole procedure is summarized in Algorithm 2.

Step 1: Update E With W, V and Z fixed, the problem becomes:

$$\min_E \frac{1}{2}\|E - G\|_F^2 + \frac{1}{\mu}\|E\|_{2,1}, \tag{15}$$

where $G = X - XW - Z - \frac{1}{\mu}\Lambda$. Note that Eq.(15) can be decoupled as:

$$\min_{\mathbf{e}_i} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{e}_i - \mathbf{g}_i\|_2^2 + \frac{1}{\mu}\|\mathbf{e}_i\|_2. \tag{16}$$

This problem can be efficiently solved by the soft-thresholding operator (Bach *et al.* 2012) with following closed form solution:

$$\mathbf{e}_i = \begin{cases} (1 - \frac{1/\mu}{\|\mathbf{g}_i\|_2})\mathbf{g}_i, & \|\mathbf{g}_i\|_2 > \frac{1}{\mu} \\ \mathbf{0}, & \|\mathbf{g}_i\|_2 \leq \frac{1}{\mu} \end{cases} \tag{17}$$

Step 2: Update V With W, E and Z fixed, the problem becomes:

$$\min_{\|V\|_{2,0}=k} \|V - Q\|_F^2, \tag{18}$$

where $Q = W + \frac{1}{\mu}\Sigma$. Let Ind be the index set corresponding to the first k biggest $\|\mathbf{q}_i\|_2$, then the optimal solution is:

$$\mathbf{v}_i = \begin{cases} \mathbf{q}_i, & i \in Ind \\ \mathbf{0}, & i \notin Ind \end{cases} \tag{19}$$

Step 3: Update W With E, V and Z fixed, the problem becomes:

$$\begin{aligned} \min_W \frac{\mu}{2}\|E - X + XW + Z + \frac{1}{\mu}\Lambda\|_F^2 \\ + \frac{\mu}{2}\|W - V + \frac{1}{\mu}\Sigma\|_F^2 + \alpha Tr(W^T X^T L X W) \end{aligned} \tag{20}$$

Take derivative with respect to W and set it to zero, we get

$$W = P^{-1}\left(X^T(X - E - Z - \frac{1}{\mu}\Lambda) + V - \frac{1}{\mu}\Sigma\right), \tag{21}$$

where $P = (X^T \hat{L} X + I_d)$, and $\hat{L} = \frac{2\lambda}{\mu}L + I_n$ is a sparse and symmetric matrix.

Step 4: Update Z With W, V and E fixed, the problem becomes:

$$\min_Z \frac{\mu}{2}\|H + Z\|_F^2 + \beta\|Z\|_F^2, \tag{22}$$

where $H = E - X + XW + \frac{1}{\mu}\Lambda$. Taking derivative with respect to Z and setting it to zero, we have

$$Z = -\frac{\mu}{\mu + 2\beta}H. \tag{23}$$

Since Eq.(10) is a non-convex problem, Algorithm 2 will find a local solution. The convergence of ALM method has been discussed in previous works such as (Bertsekas 1982). The overall computational complexity of the proposed method in each iteration is about $O(d^3 + nd^2)$, which is mainly the cost of calculating the selection matrix W. Note that the inverse of matrix P in Eq.(21) can be calculated before we start the iteration process, since it only depends on the input data. Moreover, we can resort to the Woodbury formula to transform the inverse operation of $d \times d$ matrix to $n \times n$ matrix, when d is much larger than n.

Algorithm 2 The Proposed Method RRCS

```
Input: Data matrix $X \in \mathbb{R}^{n \times d}$, trade-off parameters $\alpha$,$\beta$,
    and the number of selected features $k$.
Initialize $W = I_d$, $Z = 0 \in \mathbb{R}^{n \times d}$, initialize $\Lambda \in \mathbb{R}^{n \times d}$
    and $\Sigma \in \mathbb{R}^{d \times d}$ randomly. Initialize $\mu = 0.1$, $\rho = 1.01$.
while not converge do
    Update $E$ by Eq.(17).
    Update $V$ by Eq.(19).
    Update $W$ by Eq.(21).
    Update $Z$ by Eq.(23).
    Update $\Lambda$ by $\Lambda := \Lambda + \mu(E - X + XW + Z)$.
    Update $\Sigma$ by $\Sigma := \Sigma + \mu(W - V)$.
    Update $\mu$ by $\mu := \rho\mu$.
end while
Output: The index set $Ind$ of the selected features.
```

Experiments

Datasets

We use eight widely used benchmark datasets in our experiments. There are three microarray datasets: TOX[1], CLL-SUB[1] and MLL[2], three face datasets: JAFFE[3], ORL[4] and Yale[4], one object dataset COIL-20[4] and one shape dataset MPEG-7[5]. We provide a brief description of these datasets below.

TOX and CLL-SUB are from the GEO gene expression data depository with retrieval ID GDS1454 and GDS968, respectively. MLL contains 72 samples of three classes: acute lymphoblastic leukaemia, acute myeloid leukaemia and mixed-lineage leukaemia. JAFFE contains images of facial expressions posed by ten Japanese female models. The images in ORL were taken against a dark homogeneous background with the subjects in an upright, frontal position. YALE contains grayscale images under variable illuminations. COIL-20 is an object dataset with the images captured from varying angles. In MPEG-7, the shape classes are very distinct, but the dataset shows substantial within-class variations. All datasets are standardized to zero-mean and normalized by standard deviation. The statistics are summarized in Table 1.

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

[2] http://www.escience.cn/people/fpnie/papers.html

[3] http://www.kasrl.org/jaffe.html

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

[5] http://www.dabi.temple.edu/∼shape/MPEG7/dataset.html

Table 1: Description of Datasets

Dataset	Type	# Samples	# Dim	# Classes
TOX	gene	171	5748	4
CLL-SUB	gene	111	11340	3
MLL	gene	72	12582	3
JAFFE	face	213	1024	10
ORL	face	400	1024	40
YALE	face	165	1024	15
COIL-20	object	1440	1024	20
MPEG-7	shape	4096	1400	70

Comparison methods

We compare the proposed RRCS with the following state-of-the-art or specifically designed unsupervised feature selection methods.

- AllFea is taken as the baseline method which selects all the features.
- JELSR(Hou *et al.* 2011) selects features via joint embedding learning and sparse regression.
- UDFS(Yang *et al.* 2011) selects the most discriminative feature subset from the whole feature set in batch mode.
- NDFS(Li *et al.* 2012) has similar idea with JELSR but imposes a nonnegative constraint on the indicator matrix.
- RUFS(Qian and Zhai 2013) incorporates robust nonnegative matrix factorization, local learning and robust feature selection into a joint framework.
- RSR(Zhu *et al.* 2015) selects features by solving the joint $\ell_{2,1}$-norm minimization problem with self-representation.
- RSR-E is an extended version of RSR which explicitly considers the representation residue in the formulation. Concretely, RSR-E aims to solve the problem of $\min_{W,Z} \|X - XW - Z\|_{2,1} + \lambda\|W\|_{2,1} + \beta\|Z\|_F^2$.
- RRCS-S is a simplified version of our method RRCS. It does not consider the representation residue and does not add the graph regularization term. The objective function is $\min_W \|X - XW\|_{2,1} s.t. \|W\|_{2,0} = k$.
- AUFS(Qian and Zhai 2015) selects features via minimizing joint adaptive loss and $\ell_{2,0}$-norm regularization.
- SOGFS(Nie *et al.* 2016) performs feature selection and local structure learning simultaneously and thus the similarity matrix can be determined adaptively.

Experimental setup

The trade-off parameters of each compared method are tuned by grid search from the value set used in each paper, and the nearest neighbor parameter is fixed to be 5 when we construct the graph Laplacian matrix. For our method RRCS, we tune the parameters α and β from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$. We report the best results from the optimal parameters, and note that different parameters may be used for different datasets. Following previous works, we set the number of selected features k as $\{50, 100, 150, 200, 250, 300\}$ for all datasets.

For the selected features, K-means algorithm is repeated 20 times with random initialization to evaluate the average clustering performance. It is reasonable to assume that the better the selected features are, the higher results will be obtained. The clustering quality is measured by Normalized Mutual Information (NMI) and Accuracy (Acc)(Cai *et al.* 2005). The values of NMI and Acc range from 0 to 1 with higher score corresponding to better performance.

Clustering accuracy is the average performance of label matching results between ground truth labels and predicted labels.Formally, Acc is defined as follows:

$$Acc = \frac{1}{n}\sum_{i=1}^{n} \delta(y_i, map(p_i)),$$

where y_i and p_i are the ground truth label and the predicted label, respectively. $\delta(x, y)$ equals to 1 if $x = y$ and equals to zero otherwise. $map(\cdot)$ is the permutation mapping function that maps the predicted label to different true labels. The maximal fraction is taken as the final clustering accuracy.

Let C, C' denote the set of true clusters and predicted clusters, respectively. NMI is defined as:

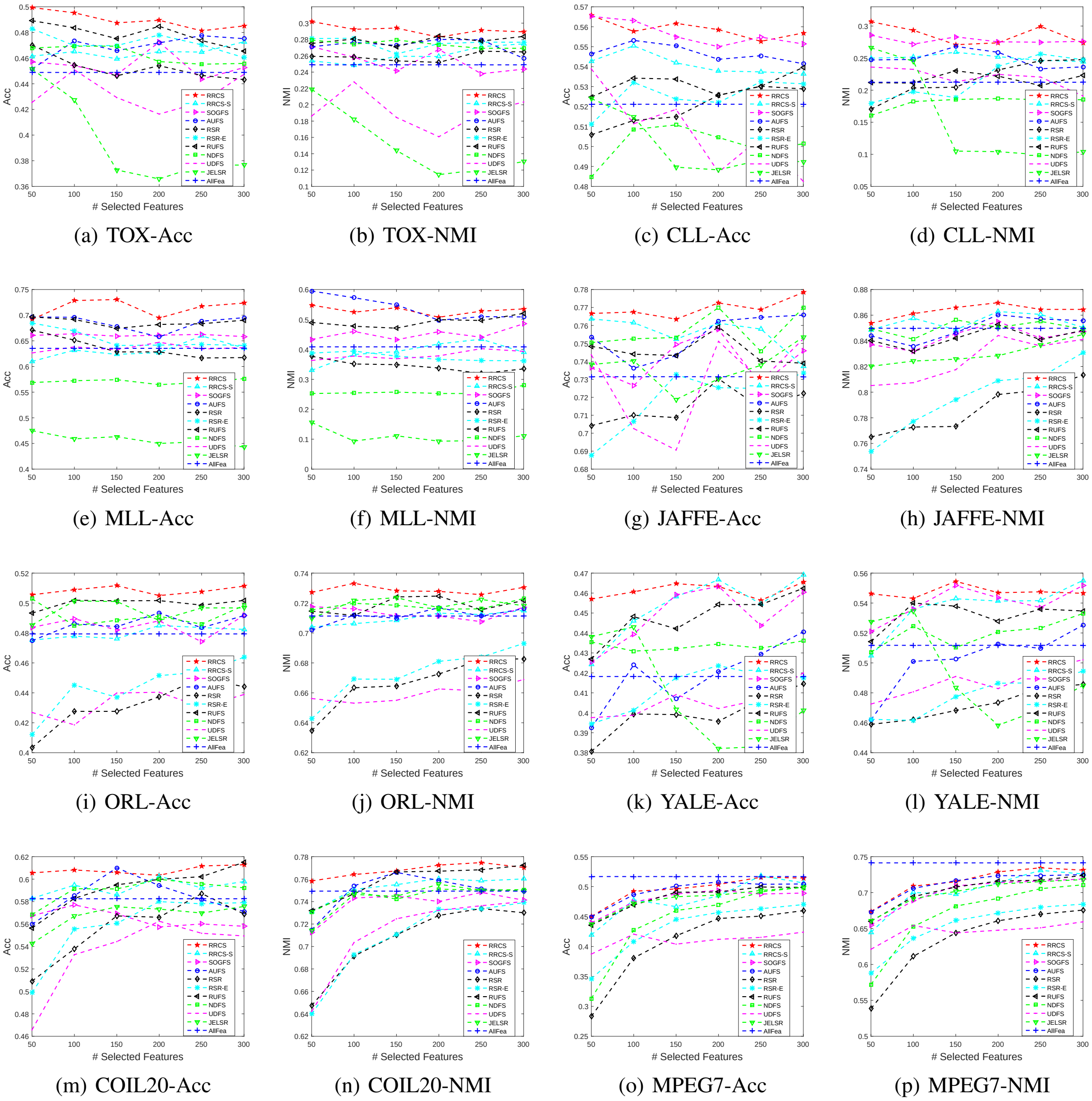
$$NMI = \frac{\sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot log_2 \left(\frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}\right)}{max\left(H(C), H(C')\right)},$$

where $p(c_i)$ and $p(c_j)$ are the probability that a data point belongs to the cluster c_i and c_j, respectively. $p(c_i, c_j)$ is the joint probability and $H(\cdot)$ is the entropy. We can see that $NMI = 1$ if the two clusters are identical, and $NMI = 0$ if they are independent.

Experimental Results and Analysis

In terms of Acc and NMI, the clustering results of different methods on the eight datasets are reported in Figure 2. We have the following observations and analysis:

- Generally speaking, the clustering performance on the selected features is better than the performance on all features. Perhaps it is because many noise features will be brought in when we select more features. The results demonstrate the necessity and effectiveness of feature selection, which both reduces the computational cost and improves the clustering performance.
- It tends to select fewer features on gene datasets since there is a trend of performance degradation. This is consistent with the fact that usually only a few of genes are associated with a given disease or biological function.
- We may find on dataset MPEG-7, the baseline AllFea achieves the best clustering results. The reason may be that MPEG-7 is a shape dataset with binary images and thus there is no noise in the images.
- Comparing to RSR, RSR-E achieves better performance in most of the cases, which demonstrates the effectiveness of considering the representation residue. RRCS-S also achieves better results than RSR on most of the datasets, demonstrating the superiority of using $\ell_{2,0}$-norm constraint to do feature selection. Therefore, the proposed method RRCS, which considers the representation



(a) TOX-Acc (b) TOX-NMI (c) CLL-Acc (d) CLL-NMI

(e) MLL-Acc (f) MLL-NMI (g) JAFFE-Acc (h) JAFFE-NMI

(i) ORL-Acc (j) ORL-NMI (k) YALE-Acc (l) YALE-NMI

(m) COIL20-Acc (n) COIL20-NMI (o) MPEG7-Acc (p) MPEG7-NMI

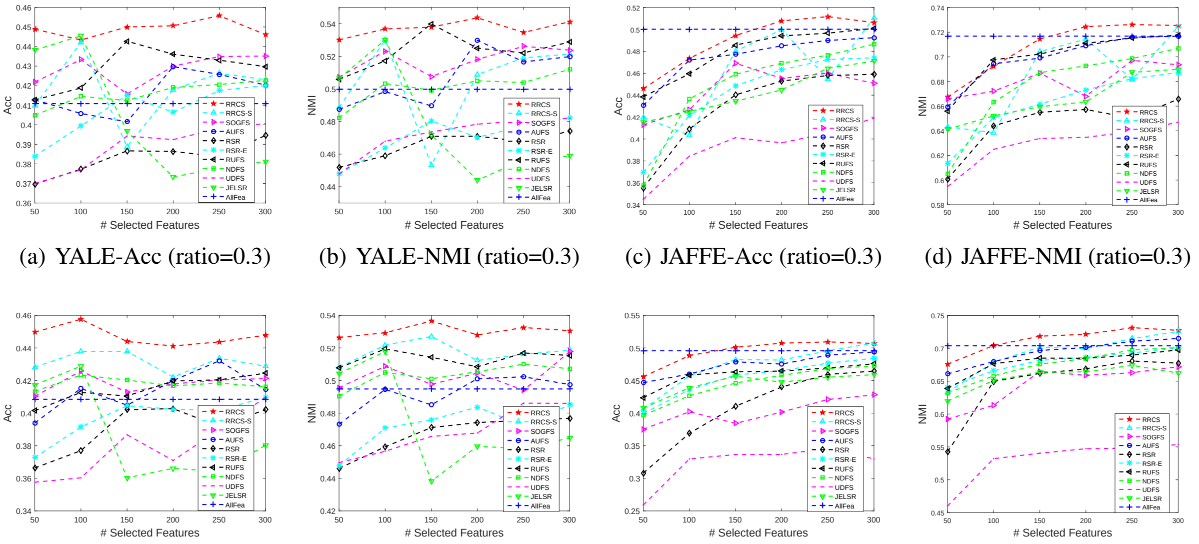
Figure 2: Clustering performance of different methods on their selected features.

residue, the local structure and the $\ell_{2,0}$-norm constraint simultaneously, achieves the best or top-3 average performance on all datasets.

- It is still an open problem to decide the optimal number k and we would better make the decision depending on the specific task and the input data in real-life applications. Since k has explicit meaning in our RRCS, thus we can avoid the burden of tuning the parameter. As a simplified version of RRCS, the method RRCS-S which also achieves competitive performance on the datasets is parameter free with a given k.

To verify the robustness of the proposed method, we conducted experiments on Yale and MPEG-7 datasets with Gaussian white noise. Concretely, 30% and 50% images were randomly selected to add the noise of mean 0 and variance 0.01. The clustering results are reported in Figure 3. We can find that the performance of baseline AllFea on MPEG-7 with explicit noise is no longer the best one, verifying the first and the third observations in Figure 2. The ratio of noisy images has smaller effect on our RRCS than on other methods, since RRCS not only uses the robust $\ell_{2,1}$-norm loss term, but also considers the representation residue which can be viewed as a modeling of the noise and outliers.

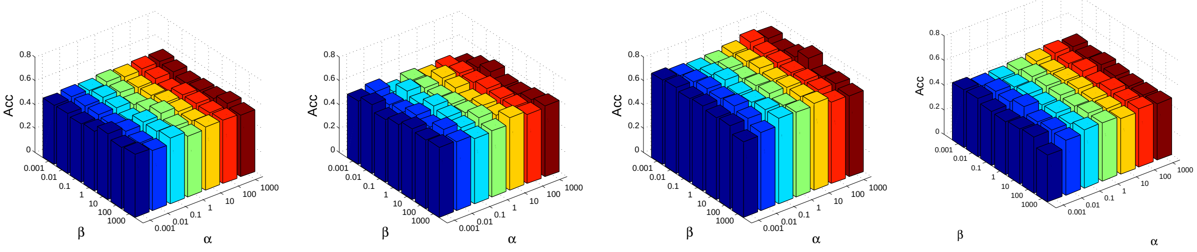
We also studied the effect of parameters α and β to the final performance on CLL-SUB, COIL-20, JAFFE and ORL. The experimental results in terms of Acc are shown in Fig-



(a) YALE-Acc (ratio=0.3) (b) YALE-NMI (ratio=0.3) (c) JAFFE-Acc (ratio=0.3) (d) JAFFE-NMI (ratio=0.3)

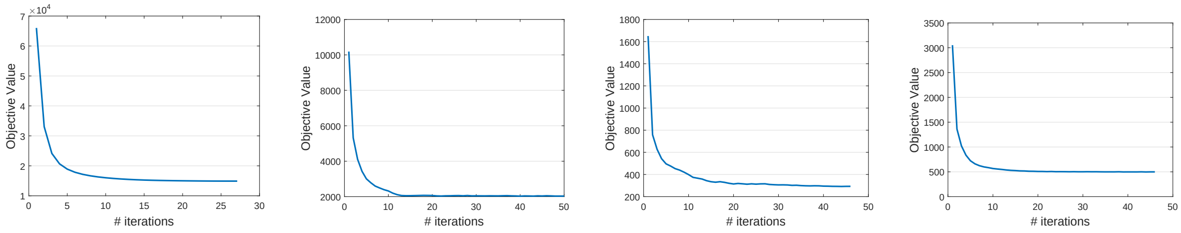
(e) YALE-Acc (ratio=0.5) (f) YALE-NMI (ratio=0.5) (g) JAFFE-Acc (ratio=0.5) (h) JAFFE-NMI (ratio=0.5)

Figure 3: Clustering performance on datasets with Gaussian white noise.



(a) CLL-SUB (b) COIL-20 (c) JAFFE (d) ORL

Figure 4: Parameter effect to clustering performance.



(a) CLL-SUB (b) COIL-20 (c) JAFFE (d) ORL

Figure 5: Curves of the objective function value.

ure 4. It can be seen that our RRCS is not very sensitive to the regularization parameters in a wide range. Figure 5 plots the curves of the objective value on the same four datasets. We can see that the proposed method has rapid convergence.

Conclusion

In this paper, we have proposed an approach called RRCS for unsupervised feature selection. RRCS is formulated from the viewpoint of self-representation with the overstrict assumption of hard linear constraint being relaxed by explicitly considering the representation residue to better deal with the nonlinear case. Feature selection is performed by solving the optimization problem with the non-smoothed $\ell_{2,1}$-norm loss term under the $\ell_{2,0}$-norm constraint. A graph regularization term is also added to preserve the local structure of data space. Extensive experiments have demonstrated the effectiveness of RRCS, comparing to state-of-the-art methods.

References

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *TPAMI*, 19(7):711–720, 1997.

Dimitri P Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics*, 1982.

Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *TKDE*, 17(12):1624–1637, 2005.

Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised discriminant analysis. In *ICCV*, pages 1–7, 2007.

Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD*, pages 333–342, 2010.

Xiao Cai, Feiping Nie, and Heng Huang. Exact top-k feature selection via $\ell_{2,0}$-norm constraint. In *IJCAI*, pages 1240–1246, 2013.

Liang Du, Zhiyong Shen, Xuan Li, Peng Zhou, and Yi Dong Shen. Local and global discriminative learning for unsupervised feature selection. In *IEEE ICDM*, pages 131–140, 2013.

Quanquan Gu, Zhenhui Li, and Jiawei Han. Joint feature selection and subspace learning. In *IJCAI*, pages 1294–1299, 2011.

Chenping Hou, Feiping Nie, Dongyun Yi, and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI*, volume 22, page 1324, 2011.

Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, Hanqing Lu, et al. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

Xinwang Liu, Lei Wang, Jian Zhang, Jianping Yin, and Huan Liu. Global and local structure preservation for feature selection. *TNNLS*, 25(6):1083–1095, 2014.

Mahdokht Masaeli, Glenn Fung, and Jennifer G. Dy. From transformation-based dimensionality reduction to feature selection. In *ICML*, pages 751–758, 2010.

Mahdokht Masaeli, Yan Yan, Ying Cui, Glenn Fung, and Jennifer G. Dy. Convex principal feature selection. In *SIAM ICDM*, pages 619–628, 2010.

Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *TIP*, 19(7):1921–1932, 2010.

Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *AAAI*, pages 1302–1308. AAAI press, 2016.

Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *IJCAI*, pages 1621–1627, 2013.

Mingjie Qian and Chengxiang Zhai. Joint adaptive loss and ℓ_2/ℓ_0-norm minimization for unsupervised feature selection. In *IJCNN*, pages 1–8. IEEE, 2015.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Lei Shi, Liang Du, and Yi-Dong Shen. Robust spectral learning for unsupervised feature selection. In *ICDM*, pages 977–982. IEEE, 2014.

Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu. Discriminant analysis for unsupervised feature selection. In *SIAM ICDM*, pages 938–946. SIAM, 2014.

Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *CVPR*, pages 586–591. IEEE, 1991.

Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, volume 22, page 1589, 2011.

M. Zhang, C. Ding, Y. Zhang, and F. Nie. Feature selection at the discrete limit. In *AAAI*, 2014.

Zheng Zhao, Lei Wang, and Huan Liu. Efficient spectral feature selection with minimum redundancy. In *AAAI*, 2010.

Zheng Zhao, Lei Wang, Huan Liu, and Jieping Ye. On similarity preserving feature selection. *TKDE*, 25(3):619–632, 2013.

Zhou Zhao, Xiaofei He, Deng Cai, Lijun Zhang, Wilfred Ng, and Yueting Zhuang. Graph regularized feature selection with data reconstruction. *TKDE*, 28(3):1–1, 2015.

Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon C. K Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.