Coupled Dictionary Learning for Unsupervised Feature Selection

Pengfei Zhu[1], **Qinghua Hu**[1], **Changqing Zhang**[1], **Wangmeng Zuo**[2]
[1]School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{zhupengfei}@tju.edu.cn

Abstract

Unsupervised feature selection (UFS) aims to reduce the time complexity and storage burden, as well as improve the generalization performance. Most existing methods convert UFS to supervised learning problem by generating labels with specific techniques (e.g., spectral analysis, matrix factorization and linear predictor). Instead, we proposed a novel coupled analysis-synthesis dictionary learning method, which is free of generating labels. The representation coefficients are used to model the cluster structure and data distribution. Specifically, the synthesis dictionary is used to reconstruct samples, while the analysis dictionary analytically codes the samples and assigns probabilities to the samples. Afterwards, the analysis dictionary is used to select features that can well preserve the data distribution. The effective $L_{2,p}$-norm ($0 < p \leq 1$) regularization is imposed on the analysis dictionary to get much sparse solution and is more effective in feature selection. We proposed an iterative reweighted least squares algorithm to solve the $L_{2,p}$-norm optimization problem and proved it can converge to a fixed point. Experiments on benchmark datasets validated the effectiveness of the proposed method.

Introduction

With the ubiquitous use of digital imaging devices, mobile terminals and social networks, mountains of high-dimensional data explosively emerge and grow. Curse of dimensionality leads to great storage burden, high time complexity and failure of the classic learning machines (Wolf and Shashua 2005). Feature selection searches the most representative and discriminative features by keeping the data properties and removing the redundancy. According to the availability of the label information, feature selection can be categorized into unsupervised (He, Cai, and Niyogi 2005), semi-supervised (Benabdeslem and Hindawi 2014), and supervised (Nie et al. 2010) ones. Because of the diverse data structure, algorithms are also developed for multi-task (Hernández-Lobato, Hernández-Lobato, and Ghahramani 2015), multi-label (Chang et al. 2014) and multi-view (Qian and Zhai 2014) feature selection.

Researchers developed many feature selection methods, including filter, wrapper, and embedding methods. Filter methods use indices that reflect data properties, e.g., variance, Fisher score, Laplacian Score (He, Cai, and Niyogi 2005), consistency (Dash and Liu 2003), to evaluate features. Different from filter methods, wapper methods rely on the learning machines. The classification or clustering performances of the learning machines are used to evaluate features (Guyon and Elisseeff 2003). For employing greedy or genetic algorithms to search a subset of features, wrapper methods are computationally intensive and therefore intractable for large-scale problems. Embedding methods perform feature selection in model construction. A feature weight vector or matrix can be learned to reflect the feature importance (Wang, Tang, and Liu 2015).

Among of these approaches, *unsupervised feature selection* (UFS) is more challenging due to the lack of label information. In unsupervised scenarios, the key factors for feature selection are locality preserving, cluster structure, and self-representation. Motivated by the intuition that nearby samples should belong to the same topic, locality preserving is widely used in feature selection, subspace learning, semi-supervised learning, etc. Laplacian Score is proposed to reflect the locality preserving power of features (He, Cai, and Niyogi 2005). In manifold learning, it is assumed that the high-dimensional data are nearly lying on a low-dimensional manifold. Hence, manifold regularization is used in unsupervised feature selection algorithms to preserve sample similarity (Li et al. 2012; Tang and Liu 2012; Wang, Tang, and Liu 2015). Similar to the class labels in supervised cases, cluster structure indicates the affiliation relations of samples, and it can be discovered by spectral clustering (SPEC (Zhao and Liu 2007), MCFS (Cai, Zhang, and He 2010), matrix factorization (NDFS (Li et al. 2012), RUFS (Qian and Zhai 2013), EUFS (Wang, Tang, and Liu 2015)) or linear predictors (UDFS (Yang et al. 2011), JELSR (Hou et al. 2011)). Due to the feature correlations, the self-representation property of features considers that one feature can be represented by a linear combination of other features. The self-representation matrix reflects the importance of features in reconstruction (Zhu et al. 2015). For different feature selection methods, the three key factors are considered individually or simultaneously.

Beyond the success in feature selection (Xiang, Tong, and Ye 2013), sparse learning has shown its power for data reconstruction (Mairal, Elad, and Sapiro 2008). For embed-

ding methods in unsupervised feature selection, matrix factorization is used to generate pseudo class labels, i.e., cluster indicators. As it is difficult to solve the discrete constraints, the constraints are relaxed to be non-negative and orthotropic (Tang and Liu 2012). The problems with matrix factorization in the existing unsupervised feature selection methods are: (1) The factorization error is large because the constraints on the cluster indicator matrix is too restrictive; (2) The learned basis matrix does not model the possible data variations well; (3) The cluster distribution should follow some data distribution priors, which cannot be reflected by matrix factorization. In fact, matrix factorization can be considered as a kind of data reconstruction. From the viewpoint of sparse representation and dictionary learning, the bases and cluster indicator in matrix factorization correspond to the dictionary and representation coefficients matrix, respectively. Similar to cluster indicator, representation coefficients can reflect data distribution as well. In image classification, an extension of the spatial pyramid matching method was developed by generalizing vector quantization to sparse coding and achieved superior performance (Yang et al. 2009). Compared with matrix factorization , dictionary learning can learn an over-complete dictionary with more possible variations. Additionally, the reconstruction error can be much smaller because of the less restrictive constraints. We can also specialize the representation coefficients with data priors to better model the data distribution.

In this paper, we propose a novel coupled dictionary learning method (CDL-FS) for unsupervised feature selection. Different from the existing methods (e.g., NDFS (Li et al. 2012), RUFS (Qian and Zhai 2013), EUFS (Wang, Tang, and Liu 2015)) that use matrix factorization to generate cluster indicators, we reconstruct the data by dictionary learning and use the representation coefficients to model the data distribution. By imposing group sparsity regularization (i.e., $L_{2,p}$-norm, $0 < p \leq 1$) on the feature weight matrix, redundancy is removed. Our main contributions include:

- A coupled analysis-synthesis dictionary learning framework is proposed for unsupervised feature selection. The synthesis dictionary is used to reconstruct the samples while the analysis dictionary analytically codes the samples. Our analysis dictionary can select the important features that can well preserve the data distribution.
- $L_{2,p}$-norm ($0 < p \leq 1$) regularization is imposed on the analysis dictionary to perform sparse feature selection and remove redundancy features. An efficient iterative reweighted least squares (IRLS) algorithm is developed with guaranteed convergence to a fixed point.
- Experiments on benchmark databases show that CDL-FS outperforms the state-of-the-art unsupervised feature selection methods. Our method is robust for different p values, especially, when $p = 0.8$, CDL-FS achieves better results in terms of both classification and clustering performance.

Problem statement

Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be the data matrix with each column $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ being a sample. $\mathcal{F} = \{f_1; ...; f_j; ...; f_d\}$ denotes the feature matrix, where f_j is the j^{th} feature vector. The objective of unsupervised feature selection is to select a subset of features from $\mathcal{F}$. Embedding methods perform feature selection by learning a feature weight matrix $\mathbf{V}$ in model construction. To generate cluster indicators, matrix factorization is introduced to cluster $\mathbf{X}$ into k clusters $\{C_1, C_2, ..., C_k\}$. The clustering model under matrix factorization framework is formulated as follow:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{A}\|_F^2 \\ &s.t. \mathbf{A} \in \{0,1\}^{k\times n}, \mathbf{A}\mathbf{1} = \mathbf{1}, \end{aligned} \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{d\times k}$ is the bases matrix, $\mathbf{A} \in \mathbb{R}^{k\times n}$ is the cluster indicator matrix, and $\mathbf{1}$ is the vector whose elements are all one. Because of the discrete constraints, it is difficult to solve the problem in Eq. (1). The constraints on $\mathbf{A}$ can be relaxed to orthogonality (Tang and Liu 2012). The clustering problem is formulated as a non-negative orthogonal matrix factorization model:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{A}\|_F^2 \\ &s.t. \mathbf{A}\mathbf{A}^T = 1, \mathbf{A} \geq 0. \end{aligned} \tag{2}$$

After obtaining the cluster indicator matrix $\mathbf{A}$ by solving Eq. (2), a matrix $\mathbf{V} \in \mathbb{R}^{\mathbf{k}\times\mathbf{d}}$ is introduced to project the data matrix $\mathbf{X}$ to the cluster indicator matrix $\mathbf{A}$ (Li et al. 2012; Qian and Zhai 2013). Different from NDFS and RUFS that use sparse regression to project $\mathbf{X}$ to $\mathbf{A}$, EUFS (Wang, Tang, and Liu 2015) directly combines sparse learning with matrix factorization. The bases matrix $\mathbf{U}$ is used for feature selection and the sparsity regularization is imposed on $\mathbf{U}$.

From the viewpoint of data reconstruction, we can use dictionary learning to replace matrix factorization:

$$\min_{\mathbf{U},\mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{A}\|_F^2 + \lambda R(\mathbf{A}), \tag{3}$$

where $\mathbf{U}$ and $\mathbf{A}$ are the dictionary and representation coefficients matrix, respectively. λ is a positive scalar constant and $R(\mathbf{A})$ is the regularization item imposed on $\mathbf{A}$.

Compared with matrix factorization, the advantages of dictionary learning are: (1) We can learn an over-complete dictionary $\mathbf{U}$, which covers more data variations.; (2) The reconstruction error can be much lower because compared with Eq. (1) and Eq. (2) there are no constraints; (3) We can specialize $\mathbf{A}$ using $R(\mathbf{A})$ (e.g., (group) sparsity regularization), which can take the data distribution priors into account. In image classification, great improvements have been achieved by extending spatial pyramid matching from vector quantization to sparse coding (Yang et al. 2009). The coding coefficients reflect data distribution and discover the cluster structure from a new perspective.

Analysis-synthesis dictionary learning

By dictionary learning, the data distribution and cluster structure can be discovered. In this section, we propose an analysis-synthesis dictionary learning model for unsupervised feature selection.

In signal representation, a signal $\mathbf{x}$ is represented over a predefined analysis dictionary $\mathbf{V}$, e.g., Gabor dictionary.

By simple inner product operations $\mathbf{V^T x}$, the representation coefficients vector $\mathbf{a}$ is easily got, i.e., $\mathbf{V^T x} = \mathbf{a}$. The coding is fast and explicit, which makes the analysis dictionary quite attractive (Elad, Milanfar, and Rubinstein 2007; Sprechmann et al. 2013). For sparse representation with a synthesis dictionary $\mathbf{U}$, a signal $\mathbf{x}$ is represented as $\mathbf{x} = \mathbf{Ua}$. For synthesis dictionary, it is easy to learn a desired dictionary and more effective in modelling local structure of images (Gu et al. 2014).

Given a synthesis dictionary $\mathbf{U} \in \mathbb{R}^{d\times k}$, the data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ is reconstructed by the synthesis dictionary, and accordingly, the representation coefficients matrix is got. Meanwhile, an analysis dictionary $\mathbf{V} \in \mathbb{R}^{d\times k}$ can also be introduced to code $\mathbf{X}$, i.e., $\mathbf{V^T X}$. Then the coupled dictionary learning model is formulated as follows:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\mathbf{X}\right\|_F^2, \tag{4}$$

To select the features that can well preserve the data distribution, the group sparsity regularization is imposed on the analysis dictionary. Then the dictionary learning model for feature selection is:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\mathbf{X}\right\|_F^2 + \tau \left\|\mathbf{V}\right\|_{2,p}^p \\ &s.t. \left\|\mathbf{u}_i\right\|_2^2 \leq 1, i = 1, 2, ..., k, \end{aligned} \tag{5}$$

where τ is a positive scalar constant and $\mathbf{u}_i$ is the i^{th} atom of the synthesis dictionary $\mathbf{U}$. The energy of the atoms $\left\|\mathbf{u}_i\right\|_2^2 \leq 1, i = 1, 2, ..., k$ is constrained to avoid trivial solution and make the solution to Eq. (5) stable (Mairal et al. 2009). $\left\|\mathbf{V}\right\|_{2,p}^p$ is $L_{2,p}$-norm of the analysis dictionary $\mathbf{V}$. $\left\|\mathbf{V}\right\|_{2,p}^p$ is defined as follows:

$$\left\|\mathbf{V}\right\|_{2,p}^p = \sum\nolimits_{i=1}^{d} \left(\sum\nolimits_{j=1}^{k} v_{ij}^2\right)^{p/2} = \sum\nolimits_{i=1}^{d} \left\|\mathbf{v}_i\right\|^p, \tag{6}$$

where $\mathbf{v}_i$ is the i^{th} row of $\mathbf{V}$, d is the number of features and k is the number of atoms in the synthesis dictionary $\mathbf{U}$.

When the value of τ increases to a certain value, most rows of $\mathbf{V}$ become zeros. The i^{th} row $\mathbf{v_i}$ of $\mathbf{V}$ is not used in coding $\mathbf{X}$ if the the elements of $\mathbf{v_i}$ are zeros. Otherwise, for the rows with non-zero elements, they play an important role in coding $\mathbf{X}$. Hence, $\left\|\mathbf{v}_i\right\|_2$ reflects the feature importance and feature selection is performed by ranking features using $\left\|\mathbf{v}_i\right\|_2$.

The p value of $L_{2,p}$ affects the sparsity of $\mathbf{V}$. When $p = 1$, $\left\|\mathbf{V}\right\|_{2,p}^p$ is the standard $L_{2,1}$-norm. When $p = 0$, $\left\|\mathbf{V}\right\|_{2,p}^p$ is the exact number of rows with non-zero elements. Hence, the sparsity on $\mathbf{V}$ increases when the value of p decreases. In supervised feature selection, the work in (Zhang et al. 2014) showed that when $p = 0$, the best results are achieved. However, in this paper, we get different observations that the minimal value of p does not lead to the best result for unsupervised feature selection. When $0 < p < 1$, compared with $L_{2,1}$-norm, a proper p value can boost the feature selection performance to some extent (refer to results in Table 2 and Table 3).

Optimization and algorithms

The model in Eq. (5) is generally non-convex. A variable matrix $\mathbf{A}$ can be introduced and the problem in Eq.(5) is relaxed as follows:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V},\mathbf{A}} \left\|\mathbf{X} - \mathbf{U}\mathbf{A}\right\|_F^2 + \mu \left\|\mathbf{A} - \mathbf{V}^T\mathbf{X}\right\|_F^2 + \tau \left\|\mathbf{V}\right\|_{2,p}^p \\ &s.t. \left\|\mathbf{u}_i\right\|_2^2 \leq 1, i = 1, 2, ..., k, \end{aligned} \tag{7}$$

where μ is a positive scalar constant. We use alteration minimization method to solve the optimization problem in Eq. (7). The synthesis dictionary $\mathbf{U}$ and analysis dictionary $\mathbf{V}$ are initialized as random matrices with unit Frobenius norm. Then we iteratively update $\mathbf{A}$ and the analysis-synthesis dictionaries $\mathbf{U}$ and $\mathbf{V}$. The detailed optimization procedures are summarized as follows:

A-subproblem: Fix $\mathbf{U}$ and $\mathbf{V}$, and update $\mathbf{A}$. We need to solve the following least squares problem :

$$\hat{\mathbf{A}} = \operatorname{argmin} \left\|\mathbf{X} - \mathbf{U}\mathbf{A}\right\|_F^2 + \mu \left\|\mathbf{A} - \mathbf{V}^T\mathbf{X}\right\|_F^2, \tag{8}$$

The closed-form solution to Eq. (8) is

$$\hat{\mathbf{A}} = \left(\mathbf{U}^T\mathbf{U} + \mu\mathbf{I}\right)^{-1} \left(\mathbf{U}^T\mathbf{X} + \mu\mathbf{V}^T\mathbf{X}\right). \tag{9}$$

After $\mathbf{A}$ is updated, we need to update $\mathbf{U}$ and $\mathbf{V}$. The optimization problem becomes:

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{A}\right\|_F^2 \ s.t. \left\|\mathbf{u}_i\right\|_2^2 \leq 1, \tag{10}$$

$$\hat{\mathbf{V}} = \arg\min_{\mathbf{V}} \mu \left\|\mathbf{A} - \mathbf{V}^T\mathbf{X}\right\|_F^2 + \tau\|\mathbf{V}\|_{2,p}^p. \tag{11}$$

U-subproblem: For the problem in Eq. (10), a variable matrix $\mathbf{H}$ can be introduced and the optimal solution of $\mathbf{U}$ can be got by Alternating Direction method of Multipliers (ADMM) (Boyd et al. 2011).

$$\hat{\mathbf{U}} = \arg\min_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{A}\right\|_F^2 s.t. \mathbf{H} = \mathbf{U} \left\|\mathbf{h}_i\right\|_2^2 \leq 1, \tag{12}$$

Then $\hat{\mathbf{U}}$ is got by the following iteration steps:

$$\begin{cases} \mathbf{U}^{t+1} = \arg\min_{\mathbf{U}} \left\|\mathbf{X} - \mathbf{U}\mathbf{A}^t\right\|_F^2 + \beta \left\|\mathbf{U} - \mathbf{H}^t + \mathbf{S}^t\right\|_F^2 \\ \mathbf{H}^{t+1} = \arg\min_{\mathbf{H}} \beta \left\|\mathbf{U}^t - \mathbf{H} + \mathbf{S}^t\right\|_F^2 s.t. \left\|\mathbf{h}_i\right\|_2^2 \leq 1 \\ \mathbf{S}^{t+1} = \mathbf{S}^t + \mathbf{U}^{t+1} - \mathbf{H}^{t+1}, \text{update } \beta \text{ if appropriate.} \end{cases} \tag{13}$$

V-subproblem: The analysis dictionary $\mathbf{V}$ is updated by solving the optimization problem in Eq. (11). In this paper, we investigate the optimization of $L_{2,p}$-norm under $0 < p \leq 1$. The problem in Eq. (11) is convex but non-smooth when $p = 1$, which has been well solved with guaranteed convergence to the optimum solution (Nie et al. 2010). Unfortunately, , when $0 < p < 1$, the problem is non-convex. Proximal gradient algorithm and rankone update algorithm were proposed to deal with the case when $0 \leq p \leq 1$ (Zhang et al. 2014). Here, we use iterative reweighted least squares (IRLS) to solve the $L_{2,p}$-norm optimization problem.

Given the current $\mathbf{V}^t$, we define diagonal weighting matrices $\mathbf{G}^t$ as:

$$g_j^t = \frac{p}{2} \left\|\mathbf{v}_j^t\right\|_2^{p-2}, \tag{14}$$

where g_j^t is the j^{th} diagonal element of $\mathbf{G}^t$ and $\mathbf{v_j^t}$ is the j^{th} row of $\mathbf{V}^t$. Then $\mathbf{V^{t+1}}$ can be updated by solving the following weighted least squares problem:

$$\begin{aligned}\mathbf{V}^{t+1} &= \arg\min_{\mathbf{V}} Q(\mathbf{V}|\mathbf{V}^t) \\ &= \arg\min_{\mathbf{V}} \left\{ \begin{array}{l} tr\left(\left(\mathbf{A}-\mathbf{V}^T\mathbf{X}\right)^T\left(\mathbf{A}-\mathbf{V}^T\mathbf{X}\right)\right) \\ +\frac{\tau}{\mu} tr\left(\mathbf{V}^T\mathbf{G}^t\mathbf{V}\right) \end{array} \right\},\end{aligned} \tag{15}$$

Let $\frac{\partial Q(\mathbf{V}|\mathbf{V}^t)}{\partial \mathbf{V}} = 0$. We get

$$\mathbf{X}(\mathbf{X}^T\mathbf{V}-\mathbf{A}^T)+\frac{\tau}{\mu}\mathbf{G}^t\mathbf{V}=0, \tag{16}$$

Then the closed-form solution of $\mathbf{V^{t+1}}$ is

$$\mathbf{V}^{t+1} = (\mathbf{X}\mathbf{X}^T+\frac{\tau}{\mu}\mathbf{G}^t)^{-1}\mathbf{X}\mathbf{A}^T, \tag{17}$$

In Eq. (17), we need to compute the inverse of $(\mathbf{X}\mathbf{X}^T + \frac{\tau}{\mu}\mathbf{G}^t)$. However, for some application, e.g., gene expression, the feature dimension is much larger than the number of samples. Hence, the complexity to compute $\mathbf{V^{t+1}}$ would be very high. According to the Woodbury matrix identity:

$$\begin{aligned}&(\mathbf{V}+\mathbf{BCD})^{-1} = \\ &\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{D}\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{V}^{-1},\end{aligned} \tag{18}$$

Then we can further get

$$\mathbf{V}^{t+1} = \mathbf{G}^{t^{-1}}\mathbf{X}\left(\mathbf{X}^T\mathbf{G}^{t^{-1}}\mathbf{X}+\frac{\tau}{\mu}\mathbf{I}\right)^{-1}\mathbf{A}^T, \tag{19}$$

When the feature dimension is larger than the number of samples, we use Eq. (19) to update $\mathbf{V^{t+1}}$. Otherwise, Eq. (17) is used.

After $\mathbf{V}$ is updated, the diagonal matrix $\mathbf{G}$ is updated by Eq. (14). To get a stable solution, a sufficiently small tolerance value is introduced by defining

$$g_j^t = \frac{p}{2\max(\left\|\mathbf{v}_j^t\right\|_2^{2-p}, \varepsilon)}. \tag{20}$$

By iteratively updating $\mathbf{V^t}$ and $\mathbf{G^t}$, the objective value of Eq. (11) monotonically decreases and guarantees to converge to a fixed point. Let $L(\mathbf{V}) = \left\|\mathbf{A}^T-\mathbf{X}^T\mathbf{V}\right\|_F^2 + \frac{\tau}{\mu}\|\mathbf{V}\|_{2,p}$. In the following, we will prove that $L(\mathbf{V})$ can be minimized by iteratively minimizing $Q(\mathbf{V}|\mathbf{V}^t)$.

Lemma 1. *$Q(\mathbf{V}|\mathbf{V}^t)$ is a surrogate function, i.e., $L(\mathbf{V}) - Q(\mathbf{V}|\mathbf{V}^t)$ attains its maximum when $\mathbf{V}=\mathbf{V}^t$.*

Proof. *Let $F(\mathbf{V}) = L(\mathbf{V}) - Q(\mathbf{V}|\mathbf{V}^t)$. We will prove that $\forall\mathbf{V}$, there is $F(\mathbf{V}^t) - F(\mathbf{V}) \geq 0$. First, $F(\mathbf{V}^t)$ can be rewritten as,*

$$\begin{aligned}F(\mathbf{V}^t) &= L(\mathbf{V}^t) - Q(\mathbf{V}^t|\mathbf{V}^t) \\ &= \frac{\tau}{\mu}\left(\left(1-\frac{p}{2}\right)\sum_{i=1}^d\left\|\mathbf{v}_j^t\right\|_2^p\right),\end{aligned} \tag{21}$$

Then we get $F(\mathbf{V}^t) - F(\mathbf{V})$

$$\begin{aligned}&F(\mathbf{V}^t) - F(\mathbf{V}) = \\ &\frac{\tau}{\mu}\sum_{j=1}^d\left(\left(1-\frac{p}{2}\right)\left\|\mathbf{v}_j^t\right\|_2^p - \left\|\mathbf{v}_j\right\|_2^p + \frac{p}{2}\frac{\left\|\mathbf{v}_j\right\|_2^2}{\left\|\mathbf{v}_j^t\right\|_2^{2-p}}\right).\end{aligned} \tag{22}$$

Algorithm 1: Algorithm of coupled dictionary learning (CDL-FS) for unsupervised feature selection

Input: $\mathbf{X}\in\mathbb{R}^{d\times n}$, μ and τ
Initialize $\mathbf{U}$ and $\mathbf{V}$.
while *not converged* **do**
  update $\mathbf{A}$ by Eq. (9);
  update $\mathbf{U}$ by solving Eq. (10) using ADMM;
  update $\mathbf{V}$ by solving Eq. (11) using IRLS;
end
Calculate feature weights $w_i = \|\mathbf{v}_i\|_2$, $i = 1, 2, ..., d$
Output: Feature weight vector $\mathbf{w}$

Let $a = \left\|\mathbf{v}_j^t\right\|_2$ and $b = \left\|\mathbf{v}_j\right\|_2$. Then we have

$$\begin{aligned}\Im_j &= \left(1-\tfrac{p}{2}\right)\left\|\mathbf{v}_j^t\right\|_2^p - \left\|\mathbf{v}_j\right\|_2^p + \tfrac{p}{2}\frac{\left\|\mathbf{v}_j\right\|_2^2}{\left\|\mathbf{v}_j^t\right\|_2^{2-p}} \\ &= \left(1-\tfrac{p}{2}\right)a^p - b^p + \tfrac{p}{2}a^{p-2}b^2\end{aligned} \tag{23}$$

$\Im_j(b)$ is a polynomial function about b. We take the first and second order derivatives of $\Im_j$ w.r.t b:

$$\Im_j^{'}(b) = p(a^{p-2}b - b^{p-1}), \tag{24}$$

$$\Im_j^{"}(b) = pa^{p-2} - (p-1)b^{p-2}, \tag{25}$$

Because $b\geq 0$, $a\geq 0$, $0<p\leq 1$, it is easy to get $\Im_j^{"}(a) = a^{p-2} > 0$, $\Im_j^{'}(a) = 0$ and $\Im_j(a) = 0$. Hence, we have $\Im_j(b)\geq\Im_j(a) = 0$ always holds. $F(\mathbf{V}^t) - F(\mathbf{V}) = \sum_{j=1}^d(\Im_j)\geq 0$, i.e., $L(\mathbf{V}) - Q(\mathbf{V}|\mathbf{V}^t)$ attains its maximum when $\mathbf{V}=\mathbf{V}^t$. □

According to the bound optimization framework, we can minimize $L(\mathbf{V})$ by iteratively minimizing the surrogate function $Q(\mathbf{V}|\mathbf{V}^t)$. Then we can get the following lemma:

Lemma 2. *Let $\mathbf{V}^{t+1} = \arg\min_{\mathbf{V}} Q(\mathbf{V}|\mathbf{V}^t)$. We have $L(\mathbf{V}^{t+1})\leq L(\mathbf{V}^t)$.*

Proof. *It is easy to see that:*

$$\begin{aligned}&L(\mathbf{V}^{t+1}) = L(\mathbf{V}^{t+1}) - Q(\mathbf{V}^{t+1}|\mathbf{V}^t) + Q(\mathbf{V}^{t+1}|\mathbf{V}^t) \\ &\left(Note: \mathbf{V}^t = \arg\max_{\mathbf{V}} L(\mathbf{V}) - Q(\mathbf{V}|\mathbf{V}^t)\right) \\ &\leq L(\mathbf{V}^t) - Q(\mathbf{V}^t|\mathbf{V}^t) + Q(\mathbf{V}^{t+1}|\mathbf{V}^t) \\ &\left(Note: \mathbf{V}^{t+1} = \arg\min_{\mathbf{V}} Q(\mathbf{V}|\mathbf{V}^t)\right) \\ &\leq L(\mathbf{V}^t) - Q(\mathbf{V}^t|\mathbf{V}^t) + Q(\mathbf{V}^t|\mathbf{V}^t) \\ &= L(\mathbf{V}^t).\end{aligned}$$

□

Hence, in each iteration, $\mathbf{V^{t+1}}$ can be updated by minimizing the surrogate function $Q(\mathbf{V}|\mathbf{V}^t)$. The proposed algorithm will finally converge to a stationary point.

In each iteration, $\mathbf{A}$ has a closed-form solution and is updated by Eq. (9). $\mathbf{U}$ is update by solving Eq. (10) and ADMM will rapidly converge. For the updating of $\mathbf{V}$, we propose to employ IRLS algorithm to solve the optimization problem of Eq. (11). The algorithm of coupled feature-selection for unsupervised feature selection is summarized in Algorithm 1.

Convergence analysis. We use alternation minimization to solve the problem in Eq. (7). The optimization of $\mathbf{A}$ is convex and $\mathbf{A}$ has a closed-form solution. For analysis-synthesis dictionary pair, the ADMM algorithm can guarantee that the optimization of $\mathbf{U}$ converges to the optimum solution. We also prove that the proposed IRLS algorithm to solve Eq. (11) can converges to a stationary point. When $p = 1$, the problem in Eq.(7) is a bi-convex problem for $\{\mathbf{A}, (\mathbf{U}, \mathbf{V})\}$. The convergence of such a problem has already been intensively studied (Gorski, Pfeuffer, and Klamroth 2007), and the proposed optimization algorithm is actually an alternate convex search (ACS) algorithm. Hence, when $p = 1$, the problem in Eq. (7) would finally converge.

We empirically find that the proposed CDL-FS algorithm converges rapidly, as shown in Figure 1. We run CDL-FS on AR face database by fixing μ and τ in Eq. (7). Additionally, different values are assigned to p and the convergence curves are similar. Figure 1 shows that when the value of p decreases, the convergence speed becomes faster. When $p = 0.4$ and $p = 0.6$, the algorithm converges in 10 iterations.

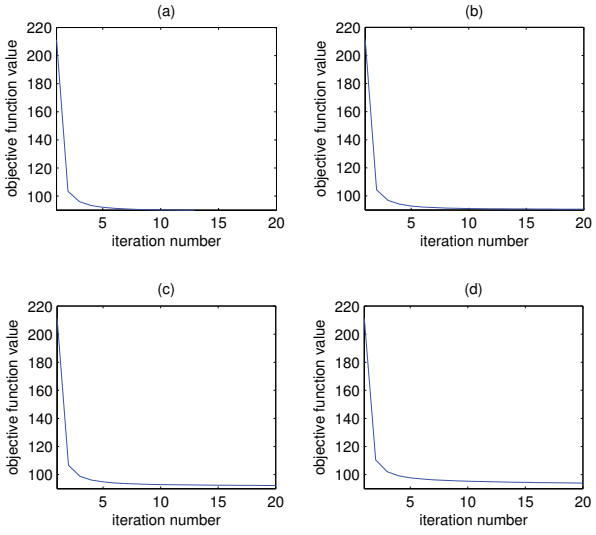

Figure 1: The convergence curve of CDL-FS with different p values. (a) p=0.4; (b) p=0.6; (c) p=0.8; (d) p=1.0.

Computational complexity. The time complexity of CDL-FS is composed of three parts, i.e., the updating of $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{V}$ in each iteration. $\mathbf{A}$ has a closed-form solution and the time complexity of updating $\mathbf{A}$ is $O(kdn + k^3 + k^2n)$, where k, d and n are the number of atoms in the synthesis dictionary $\mathbf{U}$, feature dimension and the number of samples, respectively. For the updating of $\mathbf{U}$, let T_1 be the iteration number of the ADMM algorithm. The time complexity of updating $\mathbf{U}$ is $O(T_1(dkn + k^3 + k^2d + d^2k))$. For $\mathbf{V}$, the key computation burden lies in the updating of $\mathbf{V^t}$. Hence, if $\mathbf{V^t}$ is updated by Eq. (17), the time complexity is $O(T_2(d^3 + dn^2 + dnk))$, where T_2 is the iteration number of the IRLS algorithm. If $\mathbf{V^t}$ is updated by Eq. (19), the time complexity is $O(T_2(n^3 + nd^2 + dnk))$.

Experiments

In this section, experiments are conducted to verify the effectiveness of the proposed algorithm on six benchmark datasets. The classification and clustering performance are evaluated for CDL-FS and all comparison methods. We also give the performance of CDL-FS with different p values and analyze the influence of p values.

Datasets

Six diverse publicly available datasets are selected for comparison, including one face recognition dataset (i.e., warpAR10P[1]), one handwritten digit recognition dataset (i.e., USPS[2]), one object recognition dataset (i.e., COIL20[3]), one spoken letter dataset (i.e., ISOLET[4]) and two microarray datasets (i.e., SMK-CAN-187[5]and Prostate-GE[6]). The statistics of the six datasets are shown in Table 1. The feature dimension varies between 256 and 19993 while the sample number varies between 102 and 9298.

Table 1: Summary of the benchmark datasets

DATA	Samples	Features	Classes
warpAR10P	130	2400	10
USPS	9298	256	10
COIL20	1440	1024	20
SMK-CAN-187	187	19993	2
Prostate-GE	102	5966	2
ISOLET	1560	617	26

Comparison methods

Following the common experiment setting of unsupervised feature selection, the clustering and classification performances are evaluated. We compare CDL-FS with the following representative methods:

- Laplacian Score: A filter method that selects features according to the power of locality preserving (He, Cai, and Niyogi 2005).
- SPEC: Spectral Feature Selection. SPEC is a filter method that uses spectral clustering (Zhao and Liu 2007).
- MCFS: Multi-cluster Unsupervised Feature Selection. MCFS is a filter method that uses spectral clustering and sparse regression with l_1-norm regularization (Cai, Zhang, and He 2010).
- UDFS: Unsupervised Discriminative Feature Selection. UDFS generates pseudo class labels by a linear classifier and uses $l_{2,1}$-norm regularization (Yang et al. 2011).

[1]http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html
[2]http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html
[3]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[4]https://archive.ics.uci.edu/ml/datasets/ISOLET
[5]http://featureselection.asu.edu/datasets.php
[6]https://sites.google.com/site/feipingnie/

Table 2: Classification accuracy (ACC %) of different feature selection methods. The top two results are highlighted in bold.

DATA	Laplacian	SPEC	MCFS	UDFS	RUFS	CDL-FS(0.4)	CDL-FS(0.6)	CDL-FS(0.8)	CDL-FS(1.0)
warpAR10P	70.18	76.02	73.15	85.41	84.23	88.75	89.08	**89.37**	**90.92**
USPS	87.13	53.04	87.76	89.82	92.11	**92.83**	**92.91**	92.79	91.74
COIL20	76.65	33.76	85.34	89.41	90.18	91.99	**92.83**	**92.15**	88.45
Prostate-GE	67.46	73.06	76.63	75.53	74.94	**78.46**	76.2	**76.87**	75.27
SMK-CAN-187	56.62	61.27	63.32	62.84	63.72	63.88	**64.79**	**65.59**	64.79
ISOLET	69	61	67.65	73.2	75.24	76.59	76.55	**76.97**	**76.75**

Table 3: Clustering performance (NMI %) of different feature selection methods. The top two results are highlighted in bold.

DATA	Laplacian	SPEC	MCFS	UDFS	RUFS	CDL-FS(0.4)	CDL-FS(0.6)	CDL-FS(0.8)	CDL-FS(1.0)
warpAR10P	36.26	45.74	18.17	38.72	39.86	39.22	38.37	**42.05**	**48.26**
USPS	54.73	30.14	55.89	56.95	58.48	57.97	**58.55**	58.01	**59.12**
COIL20	62.21	42.06	67.21	68.23	**70.61**	70.44	70.14	**71.42**	68.02
SMK-CAN-187	0.15	1.76	0.23	3.23	6.25	2.27	3.42	**7.27**	**6.85**
Prostate-GE	1.03	2.37	2.02	5.23	5.51	**5.94**	4.57	**5.75**	5.71
ISOLET	66.62	56.84	65.49	69.8	70.73	70.85	**71.37**	**71.23**	71.2

- RUFS: Robust Unsupervised Feature Selection. RUFS generates cluster labels by nonnegative matrix factorization and combines manifold learning (Qian and Zhai 2013).

Parameter setting

Following the experiment setting in (Yang et al. 2011; Qian and Zhai 2013), for all the compression methods, including Laplacian Score, SPEC, MCFS (Cai, Zhang, and He 2010), UDFS (Yang et al. 2011) and RUFS (Qian and Zhai 2013), the neighborhood size is set to 5 for all the six datasets. To conduct a fair comparison, for all the unsupervised feature selections methods, we tune parameters using a grid-search strategy from $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{-1}, 10^{0}, 10^{1}, 10^{3}, 10^{6}, 10^{9}\}$. For the proposed method, there are two parameters in Eq. (7), i.e., μ and τ. In the experiment, μ is fixed to 1 and τ is tuned by the grid-search strategy. Additionally, the number of atoms in the synthesis dictionary is fixed as half the number of samples. The K-means algorithm is used to evaluate the clustering performance of all the feature selection methods. As the performance of K-means is sensitive to the initialization, we run the algorithms 20 times and the average results are reported. Additionally, K nearest neighbor classifier is selected as the classifier and K is set as 1. For feature dimensions, the numbers of features are set as $\{10, 20, ..., 150\}$. As it is hard to select the feature dimension that achieves the best classification and clustering performance, we report the average results of different feature dimensions.

Experimental results

The classification and clustering results on six datasets for all the comparison methods are listed in Table 2 and Table 3. Following the experiment setting in (Yang et al. 2011; Qian and Zhai 2013), we use classification accuracy (ACC) of nearest neighbor classifier and normalized mutual information (NMI) to evaluate the performance of different feature selection methods. From the experimental results, the following observations are induced:

- CDL-FS outperforms the state-of-the-art comparison methods in terms of both clustering and classification performance. There are three reasons: first, CDL-FS uses the representation coefficients, rather than cluster labels, to reflect the cluster structure and data distribution; second, compared with matrix factorization methods, the learned synthesis dictionary can better reconstruct the data and smaller reconstruction error can be obtained; third, $L_{2,p}$-norm regularization introduces more sparsity.
- For different p values, the best average results are achieved when $p = 0.8$. A smaller value of p leads to more sparsity. Besides, the value of the regularization λ (note that μ is fixes as 1 in the experiment) also affects the sparsity of $\mathbf{V}$. The results shows that the sparsity can help to boost the performance of unsupervised feature selection to some extent.
- Although CDL-FS does not consider the locality preservation or manifold structure, our method still achieves comparable performance.

Conclusions

In this paper we proposed a novel coupled dictionary learning (CDL-FS) method for unsupervised feature selection. CDL-FS employs representation coefficients rather than cluster indicators to model the data distribution. The synthesis dictionary is used to reconstruct the data while the analysis dictionary analytically codes the data. The analysis dictionary is employed to select the features that can well preserve the data distribution. L_2, p $(0 < p \leq 1)$ -norm regularization is imposed on the analysis dictionary to introduce more sparsity. We developed an IRLS algorithm with guaranteed convergence to solve L_2, p-norm optimization problem. Experiments showed that CDL-FS achieved superior performance to the state-of-the-art methods. Additionally, the results showed that when $p = 0.8$ the best average classification and clustering performances were achieved. Therefore, a proper p value can boost the performance compared with the standard $L_{2,1}$-norm. In the future, we will take locality preserving and manifold structure into account.

Acknowledgement

This work was supported by the National Program on Key Basic Research Project under Grant 2013CB329304, the National Natural Science Foundation of China under Grants 61502332, 61432011, 61222210.

References

Benabdeslem, K., and Hindawi, M. 2014. Efficient semi-supervised feature selection: Constraint, relevance, and redundancy. *Knowledge and Data Engineering, IEEE Transactions on* 26(5):1131–1143.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1):1–122.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 333–342.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *The Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Dash, M., and Liu, H. 2003. Consistency-based search in feature selection. *Artificial intelligence* 151(1):155–176.

Elad, M.; Milanfar, P.; and Rubinstein, R. 2007. Analysis versus synthesis in signal priors. *Inverse problems* 23(3):947.

Gorski, J.; Pfeuffer, F.; and Klamroth, K. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66(3):373–407.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *Advances in Neural Information Processing Systems*, 793–801.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3:1157–1182.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *Advances in neural information processing systems*, 507–514.

Hernández-Lobato, D.; Hernández-Lobato, J. M.; and Ghahramani, Z. 2015. A probabilistic model for dirty multi-task feature selection. In *Proceedings of The 32nd International Conference on Machine Learning*, 1073–1082.

Hou, C.; Nie, F.; Yi, D.; and Wu, Y. 2011. Feature selection via joint embedding learning and sparse regression. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, 1324.

Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; and Lu, H. 2012. Unsupervised feature selection using nonnegative spectral analysis. In *The Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 689–696.

Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *Image Processing, IEEE Transactions on* 17(1):53–69.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Qian, M., and Zhai, C. 2013. Robust unsupervised feature selection. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 1621–1627.

Qian, M., and Zhai, C. 2014. Unsupervised feature selection for multi-view clustering on text-image web news data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1963–1966.

Sprechmann, P.; Litman, R.; Yakar, T. B.; Bronstein, A. M.; and Sapiro, G. 2013. Supervised sparse analysis and synthesis operators. In *Advances in Neural Information Processing Systems*, 908–916.

Tang, J., and Liu, H. 2012. Unsupervised feature selection for linked social media data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 904–912.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wolf, L., and Shashua, A. 2005. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *The Journal of Machine Learning Research* 6:1855–1887.

Xiang, S.; Tong, X.; and Ye, J. 2013. Efficient sparse group feature selection via nonconvex optimization. In *Proceedings of The 30th International Conference on Machine Learning*, 284–292.

Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, 1794–1801.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 1589–1594.

Zhang, M.; Ding, C.; Zhang, Y.; and Nie, F. 2014. Feature selection at the discrete limit. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, 1151–1157.

Zhu, P.; Zuo, W.; Zhang, L.; Hu, Q.; and Shiu, S. C. 2015. Unsupervised feature selection by regularized self-representation. *Pattern Recognition* 48(2):438–446.