Unsupervised Simultaneous Orthogonal Basis Clustering Feature Selection

Dongyoon Han and Junmo Kim
School of Electrical Engineering, KAIST, South Korea

In this paper, we propose a novel unsupervised feature selection method: Simultaneous Orthogonal basis Clustering Feature Selection (SOCFS). To perform feature selection on unlabeled data effectively, a regularized regression-based formulation with a new type of target matrix is designed. The target matrix captures latent cluster centers of the projected data points by performing the orthogonal basis clustering, and then guides the projection matrix to select discriminative features. Unlike the recent unsupervised feature selection methods, SOCFS does not explicitly use the pre-computed local structure information for data points represented as additional terms of their objective functions, but directly computes latent cluster information by the target matrix conducting orthogonal basis clustering in a single unified term of the proposed objective function.

Since the target matrix is put in a single unified term for regression of the proposed objective function, feature selection and clustering are simultaneously performed. In this way, the projection matrix for feature selection is more properly computed by the estimated latent cluster centers of the projected data points. To the best of our knowledge, this is the first valid formulation to consider feature selection and clustering together in a single unified term of the objective function. The proposed objective function has fewer parameters to tune and does not require complicated optimization tools so just a simple optimization algorithm is sufficient. Substantial experiments are performed on several publicly available real world datasets, which shows that SOCFS outperforms various unsupervised feature selection methods and that latent cluster information by the target matrix is effective for regularized regression-based feature selection.

Problem Formulation: Given training data, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ denote the data matrix with n instances where dimension is d and $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_n] \in \mathbb{R}^{m \times n}$ denote the corresponding target matrix where dimension is m. We start from the regularized regression-based formulation to select maximum r features is $\min_{\mathbf{W}} \|\mathbf{W}^T\mathbf{X} - \mathbf{T}\|_F^2 \quad s.t. \quad \|\mathbf{W}\|_{2,0} \leq r$. To exploit such formulation on unlabeled data more effectively, it is crucial for the target matrix $\mathbf{T}$ to have discriminative destinations for projected clusters. To this end, a new type of target matrix $\mathbf{T}$ is proposed to conduct clustering directly on the projected data points $\mathbf{W}^T\mathbf{X}$. We allow extra degrees of freedom to $\mathbf{T}$ by decomposing it into two other matrices $\mathbf{B} \in \mathbb{R}^{m \times c}$ and $\mathbf{E} \in \mathbb{R}^{n \times c}$ as $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ with additional constraints as

$$\min_{\mathbf{W},\mathbf{B},\mathbf{E}} \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \quad s.t.\ \mathbf{B}^T\mathbf{B} = \mathbf{I},\ \mathbf{E}^T\mathbf{E} = \mathbf{I},\ \mathbf{E} \geq \mathbf{0}, \quad (1)$$

where $\lambda > 0$ is a weighting parameter for the relaxed regularizer $\|\mathbf{W}\|_{2,1}$ that induces row sparsity of the projection matrix $\mathbf{W}$. The meanings of the constraints $\mathbf{B}^T\mathbf{B} = \mathbf{I}, \mathbf{E}^T\mathbf{E} = \mathbf{I}, \mathbf{E} \geq \mathbf{0}$ are as follows: 1) the orthogonal constraint of $\mathbf{B}$ lets each column of $\mathbf{B}$ be independent; 2) the orthogonal and the nonnegative constraint of $\mathbf{E}$ make each row of $\mathbf{E}$ has only one non-zero element [2]. From 1) and 2), we can clearly interpret $\mathbf{B}$ as the basis matrix, which has orthogonality and $\mathbf{E}$ as the encoding matrix, where the non-zero element of each column of $\mathbf{E}^T$ selects one column in $\mathbf{B}$.

While optimizing problem (1), $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ acts like clustering of projected data points $\mathbf{W}^T\mathbf{X}$ with orthogonal basis $\mathbf{B}$ and encoder $\mathbf{E}$, so $\mathbf{T}$ can estimate latent cluster centers of the $\mathbf{W}^T\mathbf{X}$. Then, $\mathbf{W}$ successively projects $\mathbf{X}$ close to corresponding latent cluster centers, which are estimated by $\mathbf{T}$. Note that the orthogonal constraint of $\mathbf{B}$ makes each projected cluster in $\mathbf{W}^T\mathbf{X}$ be separated (independent of each other), and it helps $\mathbf{W}$ to be a better projection matrix for selecting more discriminative features. If the clustering is directly performed on $\mathbf{X}$ not on $\mathbf{W}^T\mathbf{X}$, the orthogonal constraint of $\mathbf{B}$ extremely restricts the degree of freedom of $\mathbf{B}$. However, since features are selected by $\mathbf{W}$ and the clustering is carried out on $\mathbf{W}^T\mathbf{X}$ in our formulation, so the orthogonal constraint of $\mathbf{B}$ is highly reasonable. A schematic illustration of the proposed method is shown in Figure 1.

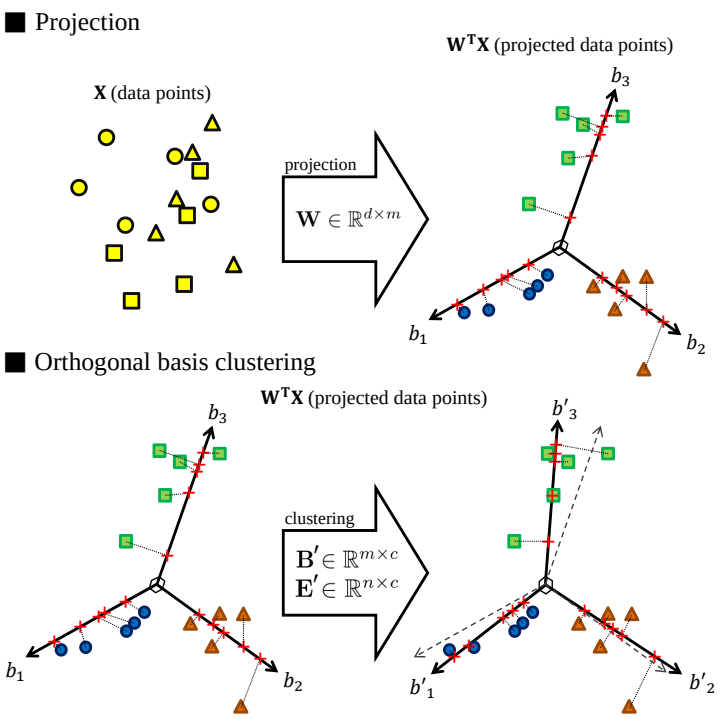

Figure 1: Schematic illustration of the proposed method. First row illustrates the projection step that maps the data points to the target matrix. Second row illustrates the orthogonal basis clustering step to discriminate latent cluster centers of the projected data points. These two steps are simultaneously conducted to select discriminative features without label information.

Optimization: Our objective function of SOCFS by rewriting problem (1) with an auxiliary variable $\mathbf{F}$ and an additional constraint $\mathbf{F} = \mathbf{E}$ is as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{E},\mathbf{F}} \ & \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} + \gamma\|\mathbf{F} - \mathbf{E}\|_F^2 \\ s.t.\ & \mathbf{B}^T\mathbf{B} = \mathbf{I},\ \mathbf{E}^T\mathbf{E} = \mathbf{I},\ \mathbf{F} \geq \mathbf{0}, \end{aligned} \quad (2)$$

where $\gamma > 0$ is another parameter to control the degree of equivalence between $\mathbf{F}$ and $\mathbf{E}$. Then an iterative optimization algorithm is proposed to solve this problem.

Experiments: We follow the same experimental setups of the previous works [1, 3, 4, 5, 6]. The experiments were conducted on seven publicly available datasets: LUNG, COIL20, Isolet1, USPS, YaleB, UMIST, AT&T. On each dataset, SOCFS is compared to the unsupervised feature selection methods [1, 3, 4, 5, 6] including all features case. It is shown that SOCFS has the best clustering results for all datasets, which indicates SOCFS selects the most discriminative features under multi-label condition. Please refer to the paper for detailed experimental results and discussions.

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

[2] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610, 2005.

[3] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *NIPS*, pages 507–514, 2005.

[4] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, pages 1026–1032, 2012.

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

[6] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.