Unsupervised Feature Selection

Yale Chang

1 Introduction

2 Spectral Feature Selection

Spectral feature selection [1] identifies relevant features by measuring their capability of preserving sample similarity.

3 Spectral Feature Selection with Minimum Redundancy [2]

This is an embedded model that evaluates the utility of a set of features jointly and can effectively remove redundant features. The algorithm is derived from a formulation based on multi-output regression and feature selection is achieved by enforcing sparsity through applying $L_{2,1}$-norm constraint on the solutions. The key idea is: to identify feature redundancy, features must be evaluated jointly.
MRSF: minimum redundancy spectral feature selection.
Given data matrix $X \in \mathbb{R}^{d\times n}$, similarity matrix $S \in \mathbb{R}^{n\times n}$, eigendecomposition gives low dimensional embedding matrix $Y \in \mathbb{R}^{n\times q}$, achieve feature selection by solving the following optimization problem

$$\min_{W,c} ||Y - X^T W||_F^2 + \lambda ||W||_{2,1} \tag{1}$$

where $W \in \mathbb{R}^{d\times q}$ is the projection matrix. [3] proposed a similar formulation

$$\min_{W} ||Y - X^T W||_{2,1} + \lambda ||W||_{2,1} \tag{2}$$

4 Joint Feature Selection and Subspace Learning [4]

$$\min_{W\in\mathbb{R}^{d\times q}} \mathrm{Tr}(W^T X L X^T W) + \lambda ||W||_{2,1} \quad s.t. \quad W^T X D X^T W = I \tag{3}$$

5 Feature Selection via Joint Embedding Learning and Sparse Regression [5]

$$\min_{W,Y} \mathrm{Tr}(Y^T L Y) + \beta(||X^T W - Y||_F^2 + \alpha ||W||_{2,1}) \tag{4}$$

where $Y \in \mathbb{R}^{n\times q}$, $Y^T Y = I_{q\times q}$, $L = (I_{n\times n} - S)^T (I_{n\times n} - S)$ is the graph Laplacian of Local Linearity Embedding, $W \in \mathbb{R}^{d\times q}$.

6 Unsupervised Feature Selection Using Nonnegative Spectral Analysis [6]

$$\min_{W,Y} \mathrm{Tr}(Y^T LY) + \beta(||X^T W - Y||_F^2 + \alpha||W||_{2,1}) \tag{5}$$

where $Y \in \mathbb{R}^{n\times q}$, $Y^T Y = I_{q\times q}$, $Y \geq 0$, $L = I_{n\times n} - D^{-1/2} S D^{-1/2}$, $W \in \mathbb{R}^{d\times q}$. When both nonnegative and orthogonal constraints are satisfied, there is only one element in each row of F is greater than zero and all the others are zeros. In that way, the learned F is more accurate, and more capable to provide discriminative information.

7 Unsupervised Feature Selection for Linked Social Media Data [7]

$$\begin{aligned} \min_{W} \ & \mathrm{Tr}(W^T XLX^T W) + \beta||W||_{2,1} + \alpha\mathrm{Tr}(W^T X(I_n - FF^T)W^T W) \\ s.t. \quad & W^T(XX^T + \lambda I_{n\times n})W = I_{q\times q} \end{aligned} \tag{6}$$

where $W \in \mathbb{R}^{d\times q}$, $L = D - S$ is a Laplacian matrix, $F = H(H^T H)^{-1/2}$ is the weighted social dimension indicator matrix, $H \in \mathbb{R}^{K\times n}$ is the social dimension indicator matrix, which can be obtained through modularity maximization.

8 Feature Selection by Joint Graph Sparse Coding [8]

$$\begin{aligned} \min_{B,G} \ & ||X - BG^T||_F^2 + \alpha\mathrm{Tr}(G^T LG) + \lambda||G^T||_{2,1} \\ s.t. \quad & \sum_{i=1}^{d}\sum_{j=1}^{q} b_{i,j}^2 \leq 1 \end{aligned} \tag{7}$$

9 Robust Unsupervised Feature Selection [9]

$$\min_{B,G,W} ||X - BG^T||_{2,1} + \nu\mathrm{Tr}(G^T LG) + \alpha||X^T W - G||_{2,1} + \beta||W||_{2,1} \tag{8}$$

where $G \in \mathbb{R}^{n\times q}$, $B \in \mathbb{R}^{d\times q}$, $W \in \mathbb{R}^{d\times q}$.

10 $L_{2,1}$-Norm Regularized Discriminative Method [10]

Supervised feature selection algorithms, e.g., Fisher score [11], robust regression [3], sparse multi-output regression [2] and trace ratio [12], usually select features according to labels of the training data. In unsupervised scenarios, label information is not available and a frequently used criterion is to select the features which best preserve the data similarity or manifold structure derived from the whole feature set [1,13,14]. Instead of evaluating the importance of each feature individually [1,13], feature correlation should be taken into account. While [2,14] apply spectral regression and consider feature correlation in two steps, this algorithm is a one-setp approach.

$$\min_{W^T W = I} \mathrm{Tr}(W^T M W) + \gamma ||W||_{2,1} \tag{9}$$

where M is constructed from local discriminative information.

11 Discriminant Analysis for Unsupervised Feature Selection [15]

12 Embedded Unsupervised Feature Selection [16]

References

[1] Z. Zhao and H. Liu, “Spectral feature selection for supervised and unsupervised learning,” in *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, ACM, 2007.

[2] Z. Zhao, L. Wang, and H. Liu, “Efficient spectral feature selection with minimum redundancy.,” in *AAAI*, 2010.

[3] F. Nie, H. Huang, X. Cai, and C. H. Ding, “Efficient and robust feature selection via joint 2, 1-norms minimization,” in *Advances in Neural Information Processing Systems*, pp. 1813–1821, 2010.

[4] Q. Gu, Z. Li, and J. Han, “Joint feature selection and subspace learning,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1294, 2011.

[5] C. Hou, F. Nie, D. Yi, and Y. Wu, “Feature selection via joint embedding learning and sparse regression,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1324, 2011.

[6] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, “Unsupervised feature selection using nonnegative spectral analysis.,” in *AAAI*, 2012.

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

[8] X. Zhu, X. Wu, W. Ding, and S. Zhang, “Feature selection by joint graph sparse coding,”

[9] M. Qian and C. Zhai, “Robust unsupervised feature selection,” in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 1621–1627, AAAI Press, 2013.

[10] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, “l2, 1-norm regularized discriminative feature selection for unsupervised learning.,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, p. 1589, 2011.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* John Wiley & Sons, 2012.

[12] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, “Trace ratio criterion for feature selection.,” in *AAAI*, vol. 2, pp. 671–676, 2008.

[13] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” in *Advances in neural information processing systems*, pp. 507–514, 2005.

[14] D. Cai, C. Zhang, and X. He, “Unsupervised feature selection for multi-cluster data,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 333–342, ACM, 2010.

[15] J. Tang, X. Hu, H. Gao, and H. Liu, “Discriminant analysis for unsupervised feature selection,”

[16] S. Wang, J. Tang, and H. Liu, “Embedded unsupervised feature selection,” 2015.