Semi-supervised Uncorrelated Feature Selection

Weichan Zhong, Xiaojun Chen and Feiping Nie

September 23, 2019

Semi-Supervised Uncorrelated Feature Selection

Weichan Zhong[1], **Xiaojun Chen**[1*], **Feiping Nie**[2]
[1]College of Computer Science and Software, Shenzhen University, Shenzhen 518060, P.R. China
[2]School of Computer Science and OPTIMAL, Northwestern Polytechnical University, Xi′an 710072, P. R. China
zhongweichan@email.szu.edu.cn, xjchen@szu.edu.cn, feipingnie@gmail.com

Abstract

In this paper, we propose an uncorrelated feature selection method for semi-supervised feature selection task, namely **SSUFS**. The new method extends the Rescaled Linear Square Regression by imposing an **Uncorrelated Regularization** (**UR**) to select only a small number of important features from highly correlated features. With this regularization, the new method is able to select lowly-nonlinear-correlated important features. SSUFS was compared with 5 feature selection methods on 5 datasets and the experimental results show the superior performance of the new method.

Introduction

In semi-supervised feature selection methods, $\ell_{2,1}$-norm is often used as a regularization item to enforce a simple group sparsity for feature selection (Chen et al. 2017; Yuan et al. 2018). However, in the standard $\ell_{2,1}$-norm, the correlations among different features are ignored. Therefore, the semi-supervised feature selection methods with the $\ell_{2,1}$-norm face a performance degeneration problem since they tend to select high-rank features that may be highly correlated with each other.

To solve this problem, we propose a novel semi-supervised feature selection method, namely **Semi-Supervised Uncorrelated Feature Selection** (**SSUFS**). In the new method, we first compute a correlation matrix to measure the nonlinear correlations between pair-wise features with "distance multivariance" (Böttcher et al. 2019). An **Uncorrelated Regularization** (**UR**) is proposed to avoid assigning both large weights to highly nonlinear correlated feature pairs, that is, it can be used to select nonlinear uncorrelated important features. With the new regularization, **SSUFS** is capable to select only a small number of important features from highly correlated features. Comprehensive experiments on 5 real-world datasets show the superiority of the proposed approach.

The Proposed Method

In semi-supervised learning, a data set $\mathbf{X} \in \mathbb{R}^{d\times n}$ with c classes consists of two subsets: a set of l labeled objects $\mathbf{X}_L = (\mathbf{x}_1, ..., \mathbf{x}_l)$ which are associated with class labels $\mathbf{Y}_L = \{\mathbf{y}_1, ..., \mathbf{y}_l\}^T \in \mathbb{R}^{l\times c}$, and a set of $u = n - l$ unlabeled objects $\mathbf{X}_U = (\mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u})^T$ whose labels $\mathbf{Y}_U = \{\mathbf{y}_{l+1}, ..., \mathbf{y}_{l+u}\}^T \in \mathbb{R}^{u\times c}$ are unknown. $\mathcal{F} = \{f_1, \cdots, f_d\}$ denotes d features. Recently, Chen et al. proposed a Rescaled Linear Square Regression (RLSR) model as follows (Chen et al. 2017):

$$\begin{aligned} &\min_{\mathbf{W},\Theta,\mathbf{b},\mathbf{Y}_U} \left(\left\| \mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y} \right\|_F^2 + \gamma \left\| \boldsymbol{\Theta}^{-1}\mathbf{W} \right\|_F^2 \right) \\ &s.t.\ \theta > 0, \mathbf{1}^T\theta = 1, \mathbf{Y}_U \geq 0, \mathbf{Y}_U\mathbf{1} = \mathbf{1} \end{aligned} \tag{1}$$

where $\mathbf{Y}_U$ are relaxed as continuous values in $[0, 1]$. $\mathbf{b} \in \mathbb{R}^c$ is the bias and $\gamma > 0$ is the regularized parameter to control the trade-off between the bias and variance of the estimate. $\Theta \in \mathbb{R}^{d\times d}$ is a diagonal matrix in which $\boldsymbol{\Theta}_{jj} = \theta_j^{1/2}$ $(1 \leq j \leq d)$ and $\theta_j > 0$ measures the importance of the j-th feature.

To conquer the "correlated features" problem in the standard $\ell_{2,1}$-norm, we compute the nonlinear feature correlation matrix $\mathbf{M}$ in which m_{ij} denotes the correlation coefficient between the i-th and j-th features. In this paper, we use the normalizing distance multivariance to measure the nonlinear correlation and m_{ij} is defined as (Böttcher et al. 2019)

$$m_{ij} = \sum_{l,k=1}^{n} \frac{A_{lk}^i}{\sqrt{\sum_{p,q=1}^n (A_{pq}^i)^2}} \frac{A_{lk}^j}{\sqrt{\sum_{p,q=1}^n (A_{pq}^j)^2}} \tag{2}$$

where A^i is a centred pair-wise distance matrix defined as

$$\mathbf{A}^i = (\mathbf{I} - \frac{1}{n}\mathbb{1})\mathbf{B}_i(\mathbf{I} - \frac{1}{n}\mathbb{1}) \tag{3}$$

where $\mathbb{1} = (1)_{j,k=1,...n}$ and $\mathbf{B}^i \in \mathbb{R}^{n\times n}$ is the pair-wise Euclidean distance matrix defined on f_i.

In order to select "diverse" features, we wish to make $\left(\left\|\mathbf{w}^i\right\|_2^2 - \left\|\mathbf{w}^j\right\|_2^2\right)^2$ big if m_{ij} is big. That is, we want to avoid assigning f_i and f_j large weights together. In this paper, we propose to introduce the following **Uncorrelated Regularization** term

$$\begin{aligned} &\max_{\mathbf{W},\mathbf{V}=diag(\mathbf{W}\mathbf{W}^T)} \left(\left\|\mathbf{w}^i\right\|_2^2 - \left\|\mathbf{w}^j\right\|_2^2\right)^2 m_{ij} \\ =&\frac{1}{2} \max_{\mathbf{W},\mathbf{V}=diag(\mathbf{W}\mathbf{W}^T)} \mathbf{V}^T\mathbf{L}_M\mathbf{V} \end{aligned} \tag{4}$$

[*]Xiaojun Chen is the corresponding author.

where $\mathbf{L}_M$ is the Laplacian matrix of $\mathbf{M}$ and $\mathbf{V} \in \mathbb{R}^{d \times 1}$ is defined as $v_j = \sum_{i=1}^{d} w_{ij}^2$. Note that if m_{ij} is small, the **Uncorrelated Regularization** regularization term tends to force both $\left\|\mathbf{w}^i\right\|_2^2$ and $\left\|\mathbf{w}^j\right\|_2^2$ to be small. In summary, the **Uncorrelated Regularization** regularization term is able to select lowly-nonlinear-correlated important features.

Introducing the constraint in Eq. (4) into problem (1) gives the following new problem

$$\min_{\mathbf{W},\mathbf{\Theta},\mathbf{b},\mathbf{Y}_U} \left\|\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right\|_F^2 + \gamma \left\|\mathbf{\Theta}^{-1}\mathbf{W}\right\|_F^2 - \eta \mathbf{V}^T \mathbf{L}_M \mathbf{V}$$

$$s.t.\ \theta > 0, \mathbf{1}^T\theta = 1, \mathbf{Y}_U \geq 0, \mathbf{Y}_U\mathbf{1} = \mathbf{1}, \mathbf{V} = diag(\mathbf{W}\mathbf{W}^T) \tag{5}$$

When $\mathbf{W}$, $\mathbf{\Theta}$ and $\mathbf{Y_U}$ are fixed, $\mathbf{b}$ can be computed as

$$\mathbf{b} = \frac{1}{n}(\mathbf{Y}^T\mathbf{1} - \mathbf{W}^T\mathbf{X}\mathbf{1}) \tag{6}$$

If $\mathbf{W}$, $\mathbf{Y}_U$ and $\mathbf{b}$ are fixed, according to (Chen et al. 2017), the optimal solution of θ_j is

$$\theta_j = \frac{\left\|\widetilde{\mathbf{w}}^j\right\|_2}{\sum_{j'=1}^{d} \left\|\widetilde{\mathbf{w}}^{j'}\right\|_2} \tag{7}$$

When $\mathbf{\Theta}$, $\mathbf{Y_U}$ and $\mathbf{b}$ are fixed, $\mathbf{W}$ can be solved with an iterative method:

1. Update the diagonal matrix $\mathbf{G}$ where $\mathbf{g}_{ii} = \mathbf{L}_M{}^i\mathbf{V}$.
2. Update $\mathbf{W} = (\mathbf{X}\mathbf{H}\mathbf{X}^T + \gamma\mathbf{\Theta}^{-2} - \eta\mathbf{G})^{-1}(\mathbf{X}\mathbf{H}\mathbf{Y})$.

Note that problem (5) is independent between different $l + 1 \leq i \leq l + u$, so we can solve the following problem individually for each $\mathbf{y}_i \in \mathbf{Y_U}$ with fixed Θ, $\mathbf{W}$ and $\mathbf{b}$

$$\min_{\mathbf{y}_i \geq 0, \mathbf{y}_i^T\mathbf{1}=1} \left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\right\|_2^2 \tag{8}$$

which has closed-form solutions and can be solved directly with the method in (Wang and nán 2013).

The algorithm to solve problem (5) is denoted as Semi-Supervised Uncorrelated Feature Selection (SSUFS). The convergence of SSUFS is ensured by the following theorem:

Theorem 1. *SSUFS monotonically decrease the objective function value of problem (5) in each iteration until the algorithm converges.*

Experimental Results and Analysis

5 benchmark datasets were selected from the UCI Machine Learning Repository [1] for this experiment. We compared SSUFS with 5 state-of-the-art semi-supervised feature selection methods, including sSelect (Zhao and Liu 2007), LSDF (Zhao, Lu, and H 2008), PRPC (Xu et al. 2016), RLSR (Chen et al. 2017) and DSFFS (Yuan et al. 2018). We set the regularization parameters of all methods in the same strategy to make the experiments fair enough, i.e., $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. The neighborhood parameters in LSDF were set to 10 for all datasets.

Table 1: The average accuracy results (the best two results on each dataset are highlighted in bold).

Name	BinAlpha	Colon	Segment	Isolet	Breast
sSelect	.290±.105	.579±.001	.670±.278	.498±.161	.575±.001
LSDF	.451±.097	.676±.043	.869±.079	.806±.099	**.679**±.049
PRPC	.377±.099	.634±.018	.854±.065	.767±.088	.621±.041
RLSR	.520±.036	.705±.035	**.928**±.035	.914±.041	.581±.05
DSSFS	**.521**±.039	**.711**±.019	**.928**±.035	**.915**±.041	.636±.063
SSUFS	**.607**±.053	**.815**±.041	.922±.037	**.923**±.050	**.681**±.054

The average accuracies of 6 methods on 5 datasets are reported in Table 1, in which we used 40% data as labeled data and 60% data as unlabeled data and test data. Overall, our proposed method SSUFS outperformed other methods on most datasets, especially on the *Colon* and *BinAlpha* datasets. To be specific, SSUFS achieves a greater than 14% average improvement on the *Colon* dataset, compared to the second-best method DSSFS. On the *BinAlpha* dataset, SSUFS achieves a nearly 16% average improvement compared to the second-best method DSFFS. SSUFS also achieved good performance on the rest datasets in average. This indicates that the reduction of the coefficients of highly correlated features indeed improves the performance of feature selection.

References

[Böttcher et al. 2019] Böttcher, B.; Keller-Ressel, M.; Schilling, R. L.; et al. 2019. Distance multivariance: New dependence measures for random vectors. *The Annals of Statistics* 47(5):2757–2789.

[Chen et al. 2017] Chen, X.; Yuan, G.; Nie, F.; and Huang, J. Z. 2017. Semi-supervised feature selection via rescaled linear regression. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1525–1531.

[Wang and nán 2013] Wang, W., and nán, M. A. C. 2013. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *Mathematics*.

[Xu et al. 2016] Xu, J.; Tang, B.; He, H.; and Man, H. 2016. Semisupervised feature selection based on relevance and redundancy criteria. *IEEE Transactions on Neural Networks and Learning Systems* PP(99):1–11.

[Yuan et al. 2018] Yuan, G.; Chen, X.; Wang, C.; Nie, F.; and Jing, L. 2018. Discriminative semi-supervised feature selection via rescaled least squares regression-supplement. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018.* AAAI Press.

[Zhao and Liu 2007] Zhao, Z., and Liu, H. 2007. Semi-supervised feature selection via spectral analysis. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 641–646.

[Zhao, Lu, and H 2008] Zhao, J.; Lu, K.; and H, X. 2008. Locality sensitive semi-supervised feature selection. *Neurocomputing* 71(10):1842–1849.

[1] https://archive.ics.uci.edu/ml/index.php