The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Discriminative Semi-Supervised Feature Selection via Rescaled Least Squares Regression-Supplement

Guowen Yuan,[1] Xiaojun Chen,[1*] Chen Wang,[1] Feiping Nie,[2] Liping Jing[3]
[1]College of Computer Science and Software, Shenzhen University, Shenzhen 518060, P.R. China
[2]School of Computer Science and Center for OPTIMAL, Northwestern Polytechnical University, Xi′an 710072, P. R. China
[3]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
243864952@qq.com, xjchen@szu.edu.cn, wangchen6771448@qq.com, feipingnie@gmail.com, lpjing@bjtu.edu.cn

Abstract

In this paper, we propose a Discriminative Semi-Supervised Feature Selection (DSSFS) method. In this method, a ϵ-dragging technique is introduced to the Rescaled Linear Square Regression in order to enlarge the distances between different classes. An iterative method is proposed to simultaneously learn the regression coefficients, ϵ-draggings matrix and predicting the unknown class labels. Experimental results show the superiority of DSSFS.

Introduction

With the rapid increase of data size, it is desirable to develop feature selection methods that are capable of exploiting both labeled and unlabeled data. During the past ten years, various semi-supervised feature selection methods have been proposed recently. Most semi-supervised feature selection methods are filter-based by ranking the features wherein the highly ranked features are selected and applied to a predictor (Zhao and Liu 2007; Xu et al. 2016). However, the filter-based feature selection could discard important features that are less informative by themselves but are informative when combined with other features. Ren et al. proposed a wrapper-type forward semi-supervised feature selection framework (Ren et al. 2008), which performs supervised sequential forward feature selection on both labeled and unlabeled data. However, this method is time consuming for high-dimensional data because it involves iterative feature subset searching. Embedded semi-supervised methods take feature selection as part of the training process, therefore, are superior to others in many respects. Recently, Chen et al. proposed a Rescaled Linear Square Regression (RLSR) for semi-supervised feature selection task (Chen et al. 2017).

In order to select features with discriminative power, it is often desired that the distances between data points in different classes are as large as possible after they are transformed. In this paper, we propose a fast Discriminative Semi-Supervised Feature Selection (DSSFS). In this method, a ϵ-dragging technique is introduced to the Rescaled Linear Square Regression to enlarge the distances between different classes. We propose an iterative method to optimize the new model. Experimental results on four real-life data sets show the superiority of DSSFS.

*Xiaojun Chen is the corresponding author.

Discriminative Semi-Supervised Feature Selection

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. Here, $\mathbf{y}_i \in \mathbb{R}^c (1 \le i \le l)$ is a binary vector in which $\mathbf{y}_i^j = 1$ if $\mathbf{x}_i$ belongs to the j-th class.

To measure the importances of d features, we introduce d scale factors θ in which $\theta_j > 0 (1 \le j \le d)$ measures the importances of the j-th feature. We use θ to evaluate the importances of the d features and the k most important features can be selected according the biggest k values in θ. To learn $\mathbf{\Theta}$ and $\mathbf{Y}_U$ simultaneously, we form the following problem

$$\min \left(\left\| \mathbf{X}^T \Theta \mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y} - (2\mathbf{Y} - 1) \circ \mathbf{M} \right\|_F^2 + \gamma \left\| \mathbf{W} \right\|_F^2 \right)$$
$$st.\ \mathbf{W},\ \mathbf{b},\ \theta > 0,\ \mathbf{1}^T\theta = 1,\ \mathbf{Y}_U \ge 0,\ \ \mathbf{Y}_U\mathbf{1} = \mathbf{1},\ \mathbf{M} \ge \mathbf{0} \tag{1}$$

where $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$ is a rescale matrix which is a diagonal matrix and $\Theta_{jj} = \sqrt{\theta_j}$. Let $\mathbf{E} = 2\mathbf{Y} - 1$, then $\mathbf{E} \in \mathbb{R}^{n\times c}$ is a constant matrix, in which $e_{il} = +1$ if the i-th object belongs to the l-th class, -1 if the i-th object does not belong to the l-th class, or 0 if $l \le i \le n$. $\mathbf{M}$ is a ϵ-draggings matrix to be learnt.

According to Theorem (1) in (Chen et al. 2017), problem (1) is equivalent to the following sparse problem

$$\min \left(\left\| \mathbf{X}^T \mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y} - (2\mathbf{Y} - 1) \circ \mathbf{M} \right\|_F^2 + \gamma \left\| \mathbf{W} \right\|_{2,1}^2 \right)$$
$$st.\ \mathbf{W},\ \mathbf{b},\ \theta > 0,\ \mathbf{1}^T\theta = 1,\ \mathbf{Y}_U \ge 0,\ \mathbf{Y}_U\mathbf{1} = \mathbf{1},\ \mathbf{M} \ge \mathbf{0} \tag{2}$$

where the optimal solution of θ is $\theta_j = \frac{\left\| \mathbf{w}^j \right\|_2}{\sum_{j=1}^d \left\| \mathbf{w}^j \right\|_2}$.

We define an iterative algorithm, named Discriminative Semi-Supervised Feature Selection (DSSFS) , to solve problem (2), in which $\mathbf{b}$, $\mathbf{W}$, $\mathbf{Y}_U$ and $\mathbf{M}$ are alternately updated in each iteration until convergency. If $\mathbf{b}$, $\mathbf{Y}_U$ and $\mathbf{M}$ are fixed, the optimal solution of $\mathbf{W}$ can be obtained by an iterative algorithm, in which an additional variable $\mathbf{Q}$ is introduced. The optimal solution of $\mathbf{W}$ is

$$\mathbf{W} = (\mathbf{X}\mathbf{H}\mathbf{X}^T + \gamma\mathbf{Q})^{-1}\mathbf{X}\mathbf{H}(\mathbf{Y} + (2\mathbf{Y} - 1) \circ \mathbf{M}) \quad (3)$$

where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with the j-th diagonal element as $q_{jj} = \frac{\sum_{v=1}^{d}\sqrt{\|\mathbf{w}^v\|_2^2+\epsilon}}{\sqrt{\|\mathbf{w}^j\|_2^2+\epsilon}}$. $\mathbf{W}$ and $\mathbf{Q}$ can be alternatively updated until convergence.

If $\mathbf{W}$, $\mathbf{Y_U}$ and $\mathbf{M}$ are fixed, the optimal solution of $\mathbf{b}$ is

$$\mathbf{b} = \frac{1}{n}\left[(\mathbf{Y} + (2\mathbf{Y} - 1) \circ \mathbf{M})^T\mathbf{1} - \mathbf{W}^T\mathbf{X}\mathbf{1}\right] \quad (4)$$

If $\mathbf{b}$, $\mathbf{M}$ and $\mathbf{W}$ are fixed, the optimal solution of $\mathbf{Y_U}$, the optimal solution of each $\mathbf{y}_i \in \mathbf{Y_U}$ can be individually updated as

$$\mathbf{y}_i = ((2\mathbf{m}^i + 1)^{\circ -1} \circ (\mathbf{W}^T\mathbf{x}_i + \mathbf{b} + \mathbf{m}^i) + \eta)_+ \quad (5)$$

where η can be obtained by solving $\mathbf{y}_i^T\mathbf{1} = 1$.

If $\mathbf{b}$, $\mathbf{Y_U}$ and $\mathbf{W}$ are fixed, the optimal solution of $\mathbf{M}$ is

$$\mathbf{M} = \max((2\mathbf{Y} - 1) \circ (\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}), \mathbf{0}) \quad (6)$$

Experimental Results and Analysis

To validate the effectiveness of DSSFS, we compared it with six state-of-the-art feature selection methods, including three semi-supervised feature selection methods sSelect (Zhao and Liu 2007), RLSR (Chen et al. 2017) and RRPC (Xu et al. 2016), two unsupervised feature selection method Laplacian Score (LS) (He, Cai, and Niyogi 2005) and MCFS (Cai, Zhang, and He 2010), and a supervised feature selection method RFS (Nie et al. 2010). In this experiment, we used four real-life data sets, i.e., the Binalpha, Usps, Colon and Musk_clean1 (MC) data sets. For each data set, we randomly selected 40% of samples as training examples, and the remaining examples were then used as the test data. The test data were also used as the unlabeled data for the semi-supervised feature selection algorithm. For unsupervised feature selection methods, we only used the training examples without labels. For the selected features, we trained linear SVM to predict the unlabeled data, the average accuracies and standard deviation were computed. We set the regularization parameter γ of LS, RFS, sSelect, RLSR and DSSFS as $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^2, 10^3\}$, λ of sSelect as $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$.

Table 1: The average accuracy $\pm$ standard deviation results (the best result on each data set is highlighted in bold).

$Algorithm$	Binalpha	Usps	Colon	MC
LS	0.466±0.16	0.601±0.12	0.651±0.07	0.687±0.04
MCFS	0.531±0.06	0.673±0.02	0.666±0.04	0.691±0.05
RFS	0.573±0.11	0.602±0.14	0.699±0.06	0.712±0.04
sSelect	0.407±0.14	0.64±0.08	0.566±0.04	0.623±0.04
PRPC	0.5±0.09	0.658±0.03	0.703±0.02	0.702±0.03
RLSR	0.556±0.11	0.677±0.03	0.751±0.05	0.719±0.04
DSSFS	**0.59**±0.1	**0.679**±0.03	**0.761**±0.03	**0.732**±0.03

The average accuracies of seven feature selection methods are summarized in Table 1. Overall, our proposed method DSSFS outperformed all other methods on these data sets. To be specific, DSSFS has nearly 3% improvement on the Binalpha data set, compared to the second best method RFS. We also notice that DSSFS defeats RLSR on these data sets. This indicates that the introduction of discriminative ϵ-dragging technique indeed improves the performance of feature selection.

Conclusions

This paper presents a Discriminative Semi-Supervised Feature Selection (DSSFS) method. Experimental results show the superiority of our method.

Acknowledgments

This research was supported by NSFC under Grant no. 61305059 and 61773268.

References

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 333–342.

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.

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

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

Ren, J.; Qiu, Z.; Fan, W.; Cheng, H.; and Yu, P. S. 2008. Forward semi-supervised feature selection. In *Proceedings of the 12th Pacific-Asia Conference in Knowledge Discovery and Data Mining*, 970–976. Berlin, Heidelberg: Springer Berlin Heidelberg.

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.

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.