Semi-Supervised Feature Selection with Adaptive Discriminant Analysis

Weichan Zhong,[1] Xiaojun Chen,[1]* Guowen Yuan,[1] Yiqin Li,[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
969566450@qq.com, xjchen@szu.edu.cn, gwyuan93@qq.com, 1550047250@qq.com, feipingnie@gmail.com

Abstract

In this paper, we propose a novel Adaptive Discriminant Analysis for semi-supervised feature selection, namely SADA. Instead of computing fixed similarities before performing feature selection, SADA simultaneously learns an adaptive similarity matrix $\mathbf{S}$ and a projection matrix $\mathbf{W}$ with an iterative method. In each iteration, $\mathbf{S}$ is computed from the projected distance with the learned $\mathbf{W}$ and $\mathbf{W}$ is computed with the learned $\mathbf{S}$. Therefore, SADA can learn better projection matrix $\mathbf{W}$ by weakening the effect of noise features with the adaptive similarity matrix. Experimental results on 4 data sets show the superiority of SADA compared to 5 semi-supervised feature selection methods.

Introduction

Since it is often costly to obtain labeled data, the study of ‘semi-supervised feature selection" has gained more and more attention. Recently, Chen proposed a semi-supervised feature selection method **RLSR** (Chen et al. 2017), in which a rescaled linear square regression is proposed to extend the least square regression for feature selection. Yuan et al. improved RLSR by introducing a ϵ-dragging technique in order to enlarge the distances between different classes (Yuan et al. 2018). In real applications, **multimodality** phenomena that samples in some classes form several separate clusters is often observed (Fukunaga 1990). However, existing semi-supervised feature selection methods cannot solve this problem.

To address the "multimodality" problem, we propose a new semi-supervised feature selection method, namely Semi-supervised Adaptive Discriminant Analysis (SADA). Instead of computing a fixed similarity matrix before performing feature selection, SADA learns an adaptive similarity matrix $\mathbf{S}$ and a projection matrix $\mathbf{W}$ simultaneously with an iterative method. In each iteration, $\mathbf{S}$ is computed from the projected distance with the learned $\mathbf{W}$ and $\mathbf{W}$ is computed with the learned $\mathbf{S}$. Therefore, SADA can better rank the features by weakening the affection of noise features with the adaptive similarity matrix. Experimental results on 4 data sets show the superiority of SADA in comparison to 5 semi-supervised feature selection methods.

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. Let $\mathbf{W} \in R^{d\times m}$ be a projection matrix where m is the projection dimension. Inspired by the paper (Xiaojun Chen and Huang 2018), we can learn $\mathbf{W}$ by solving the following objective function

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \sum_{i=1}^{n} \sum_{j\in\mathcal{M}_i^1 \bigcup \mathcal{M}_i^2} \left(\left\| \mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j) \right\|_2^2 + \epsilon \right)^{\frac{p}{2}} + \gamma \sum_{l=1}^{d} \sqrt{\mathbf{w}^l(\mathbf{w}^l)^T + \epsilon} \tag{1}$$

where $\mathcal{M}_i^1$ consists of k nearest neighbors of $\mathbf{x}_i$ in $\mathbf{X}_L$ and $\mathcal{M}_i^2$ consists of k nearest neighbors of $\mathbf{x}_i$ in $\mathbf{X}_U$. Specifically, if $\mathbf{x}_i$ is labeled, $\mathcal{M}_i^1$ consists of $\min\{k, nc_i\}$ nearest neighbors which are in the same class as $\mathbf{x}_i$ and nc_i is the number of objects in the class to which $\mathbf{x}_i$ belongs. The $\ell_{2,p}$ norm is used to obtain more sparser solution if we set a smaller p where $p \in (0, 2)$. ϵ is a sufficiently small constant, e.g. 10^{-10}, which is used to avoid zero denominators.

It is difficult to directly solve problem (1). In this paper, we propose to obtain $\mathbf{W}$ by solving the following problem

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \left[Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}_S\mathbf{X}^T\mathbf{W}) + \gamma Tr(\mathbf{W}^T\mathbf{Q}\mathbf{W}) \right] \tag{2}$$

where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a diagonal matrix in which

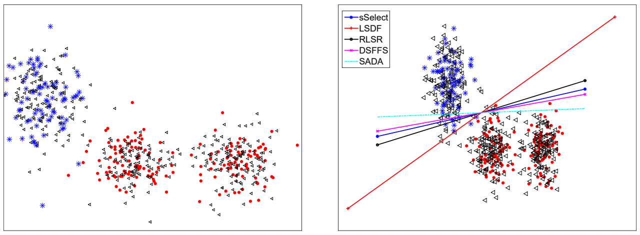
$$q_{ll} = \frac{1}{2\sqrt{\mathbf{w}^l(\mathbf{w}^l)^T + \epsilon}} \tag{3}$$

and $\mathbf{S} \in \mathbb{R}^{n\times n}$ is defined as

$$s_{ij} = \begin{cases} \frac{p}{2\left(\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 + \epsilon\right)^{1-\frac{p}{2}}} & \text{if } j \in \mathcal{M}_i^1 \bigcup \mathcal{M}_i^2 \\ 0 & otherwise \end{cases} \tag{4}$$

With fixed $\mathbf{Q}$ and $\mathbf{S}$, problem (2) can be solved directly to obtain the optimal solution to $\mathbf{W}$ as the m eigenvectors of $\mathbf{X}\mathbf{L}_S\mathbf{X}^T + \gamma\mathbf{Q}$ corresponding to the m smallest eigenvalues, where $\mathbf{L}_S = \mathbf{D}_s - \mathbf{S}$ is the Laplacian matrix of $\mathbf{S}$ and

[*]Xiaojun Chen is the corresponding author.



(a) The first two dimensions of D_1.

(b) The projection directions in the first two dimensions.

Figure 1: Projection direction results on D_1. In each figure, the blue points and red points indicate two different classes, while the black points indicate unlabeled objects.

$\mathbf{D}_s \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the i-th diagonal element as $\sum_{j=1}^{n} s_{ij}$. Then, with the new $\mathbf{W}$, we update $\mathbf{Q}$ and $\mathbf{S}$ according to Eqs. (3) and (4). Finally, $\{\left\|\mathbf{w}^j\right\|_2\}_{j=1}^{d}$ are computed from the learned $\mathbf{W}$ and the r most important features are selected out according to $\{\left\|\mathbf{w}^j\right\|_2\}_{j=1}^{d}$. The above algorithm is denoted as Semi-supervised Adaptive Discriminant Analysis (SADA). The convergence of SADA is ensured by the following theorem:

Theorem 1. *The iteration process of SADA will monotonically decrease the objective function of problem (1) in each iteration.*

Experimental Results and Analysis

We generated a synthetic data set D_1 to test the projection ability of the proposed method for feature selection. The data set consists of 12 dimensions, where the data in the first two dimensions are distributed in three Gaussian shapes while the data in the other dimensions are uniformly distributed noise features. Figure 1a shows the data set in the first two dimensions, in which two small Gaussian clusters are buried in one red class. We compared SADA with five 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). In this experiment, the projection dimension was set as 1 and the nearest neighborhoods k was set as 5. The regularization parameters in RLSR, DSFFS and SADA were set as 1 for fair comparison. The neighborhood parameters in LSDF and SADA was set to 5 for all datasets. For SADA, we set $p = 1.5$. The projection direction results are displayed in Figure 1b, which shows that if we consider separating only the red class from the blue class, the direction of projection revealed by LSDF is good. However, if we want to separate the two small classes contained within the red class, SADA achieves the best direction of projection. In this experiment, we compared six methods on four real-life data sets whose characteristics are shown in Table 1. We set parameters of all methods in the same strategy to make the experiments fair enough, i.e., $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The neighborhood parameters in LSDF and SADA was set to 10 for all datasets. p in SADA was set to 10 values from 0.1 to 1.9. The average accuracies of 6 methods on 4 datasets are reported in Table 2, in which we used 30% data as labeled data and 70% data as unlabeled data and test data. Overall, our proposed method SADA outperformed other methods on most datasets, especially on the *Colon* datasets. To be specific, SADA achieves a greater than 2% average improvement on the *Colon* dataset, compared to the second-best method PRPC. SADA also achieved good performance on the rest datasets in average. This indicates that the learnt implicit adaptive local structure learning indeed improves the performance of feature selection.

Table 1: Characteristics of 4 benchmark data sets.

Name	#Samples	#Features	#Classes
Colon	62	2000	2
Segment	2310	19	7
Srbct	63	2308	2
Glass	214	9	6

Table 2: The average accuracies of 6 semi-supervised feature selection methods on 4 benchmark data sets (the best result on each data set is highlighted in bold).

Name	Colon	Segment	Srbct	Glass
LSDF	.877±.012	.859±.089	.551±.025	**.502** ±.023
sSelect	.682±.000	.654±.295	.356±.000	.434±.082
PRPC	**.893** ±.023	.834±.060	.429±.037	.479±.019
RLSR	.841±.027	**.923** ±.032	**.591** ± .004	.492±.020
DSFFS	.841±.053	**.923** ±.046	**.593** ±.018	.492±.027
SADA	**.911** ±.019	**.909**±.044	**.591** ±.027	**.494** ±.023

References

Chen, X.; Yuan, G.; Nie, F.; and Huang, J. Z. 2017. Semi-supervised feature selection via rescaled linear regression. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1525–1531.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. San Diego, CA, USA: Academic Press Professional, Inc.

Xiaojun Chen, Guowen Yuan, W. W. F. N. X. C., and Huang, J. Z. 2018. Local Adaptive Projection Framework for Feature Selection of Labeled and Unlabeled Data. *IEEE Transactions on Neural Networks & Learning Systems* PP(99):1–12.

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, 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*. AAAI Press.

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