Late-Breaking Developments in the Field of Artificial Intelligence
Papers Presented at the Twenty-Seventh AAAI Conference on Artificial Intelligence

Discriminative Multi-Task Feature Selection

Yahong Han[1,4], Jianguang Zhang[1], Zhongwen Xu[2], Shoou-I Yu[3]
[1]School of Computer Science and Technology, Tianjin University, China
[2]College of Computer Science, Zhejiang University, China
[3]School of Computer Science, Carnegie Mellon University
[4]Tianjin Key Laboratory of Cognitive Computing and Application
{yahong,lynxzjg}@tju.edu.cn, zhongwen19901215@gmail.com, iyu@cs.cmu.edu

Abstract

The effectiveness of supervised feature selection degrades in low training data scenarios. We propose to alleviate this problem by augmenting per-task feature selection with joint feature selection over multiple tasks. Our algorithm builds on the assumption that different tasks have shared structure which could be utilized to cope with data sparsity. The proposed trace-ratio based model not only selects discriminative features for each task, but also finds features which are discriminative over all tasks. Extensive experiment on different data sets demonstrates the effectiveness of our algorithm in low training data scenarios.

Introduction

Feature selection has a two-fold role in improving both the efficiency and accuracy of data analysis (Gao et al. 2011; Nie et al. 2010; Cai et al. 2011; Zhao and Liu 2007; Yang et al. 2011). Most of the existing feature selection algorithms select features for each task independently. When we estimate models for several related tasks (Caruana 1997; Argyriou, Evgeniou, and Pontil 2008), tasks which share some common underlying representations will benefit from joint learning. Thus, we can leverage the knowledge from multiple related tasks to improve the performance of feature selection (Obozinski, Taskar, and Jordan 2006; Ma et al. 2012; Yang et al. 2013).

Most of the feature selection algorithms evaluate the importance of each feature individually and select features one by one (Duda, Hart, and Stork 2001; Tibshirani 1996; Cawley, Talbot, and Girolami 2007). A limitation is that the correlation among features is neglected. Recently, researchers have applied the $\ell_{2,1}$-norm to evaluate the importance of the selected feature jointly (Nie et al. 2010; Yang et al. 2011). More recently, researchers impose a joint regularization term on the multiple feature selection matrices (Yang et al. 2013; Han, Yang, and Zhou 2013) for better performance of feature selection. However, the discriminative information among the multiple tasks is not well exploited (Yang et al. 2013). Different from the $\ell_{2,1}$-norm used in the transfer learning (Ma et al. 2012), we use it to uncover the common irrelevant features among multiple tasks.

This paper exploits the discriminative information for multi-task feature selection. We firstly utilize the trace ratio criterion for each task to minimize the ratio of within-class scatter to the between-class scatter. An $\ell_{2,1}$-norm is imposed to each task to perform feature selection respectively. Then a joint $\ell_{2,1}$-norm is imposed so that the common irrelevant or noisy features in different tasks are uncovered.

The Objective Function and Solution

Suppose we are going to select features for t tasks. The l-th task contains m_l training data $\{x_l^i\}_{i=1}^{m_l}$ with groundtruth labels $\{y_l^i\}_{i=1}^{m_l}$ from c_l classes. We define $X_l = [x_l^1, \dots, x_l^{m_l}]$ as the data matrix of the l-th task and $Y_l = [y_l^1, \dots, y_l^{m_l}]$ as the corresponding label matrix. Given a matrix $A \in \mathbb{R}^{a\times b}$ where a and b are arbitrary numbers, $||A||_F$ is its Frobenius norm. The $\ell_{2,1}$-norm of A is defined as $||A||_{2,1} = \sum_i (\sum_j A_{ij}^2)^{\frac{1}{2}}$. In the following, $Tr(\cdot)$ represents the trace operator, I_{m_l} is the $m_l \times m_l$ identity matrix, and $\mathbf{1}_{m_l}$ is a column vector with all of its element being 1.

For the l-th task, we define the scaled category indicator matrix F_l as $F = Y_l(Y_l^T Y_l)^{-\frac{1}{2}}$. Then the between-class scatter and total class scatter for the l-th task are defined as (Duda, Hart, and Stork 2001): $S_b^{(l)} = \tilde{X}_l F_l F_l^T \tilde{X}_l^T$ and $S_{to}^{(l)} = \tilde{X}_l \tilde{X}_l^T$, where $\tilde{X}_l = X_l H_l$ and $H_l = I_{m_l} - \frac{1}{m_l}\mathbf{1}_{m_l}\mathbf{1}_{m_l}^T$ is the centering matrix. We propose the discriminative multi-task feature selection as to solve:

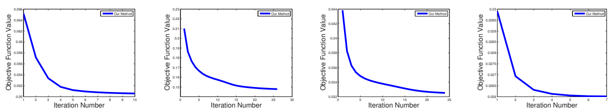
$$\min_{W_l^T W_l = I|_{l=1}^t} \sum_{l=1}^{t} \left(\frac{Tr(W_l^T \tilde{X}_l (I_{m_l} - F_l F_l^T)\tilde{X}_l^T W_l)}{Tr(W_l^T \tilde{X}_l \tilde{X}_l^T W_l)} + \lambda_1 \sum_i (\sum_j (W_{ij}^{(l)})^2)^{\frac{1}{2}} \right) + \lambda_2 \sum_i (\sum_j W_{ij}^2)^{\frac{1}{2}}, \quad (1)$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters, $W_{ij}^{(l)}$ denotes (i,j)-th element of the transformation matrix W_l for the l-th task, and $W = [W_1, \dots, W_t]$ is the jointly feature selection matrix for the t tasks. The objective function in Eq. (1) can be solved by alternatively optimizing W_l ($l = 1, \dots, t$) until convergence.

Denote $E_l = \tilde{X}_l (I_{m_l} - F_l F_l^T)\tilde{X}_l^T$ and $B_l = \tilde{X}_l \tilde{X}_l^T$ and fix W_j ($j = 1, \dots, l-1, l+1, \dots, t$), the objective function

Table 1: Classification results (Accuracy) of comparison methods.

	Full Features	Fisher Score	SBMLR	SVM-21	LSR-21	FSSI	Our Method
MIML	0.3133	0.3341	0.2469	0.3238	0.3345	0.3809	**0.3917**
USPS	0.7900	0.7939	0.5574	0.7934	0.8015	0.8031	**0.8155**
Protein	0.3812	0.3943	0.3539	0.3824	0.3876	0.4157	**0.4365**
SensIT	0.6840	0.6847	0.4397	0.7031	0.7202	0.7243	**0.7358**



(a) MIML (b) USPS (c) Protein (d) SensIT

Figure 1: Convergence curves of the objective function.

in Eq. (1) is equivalent to

$$\min_{W_l^T W_l = I} \frac{Tr(W_l^T E_l W_l)}{Tr(W_l^T B_l W_l)} + \lambda_1 Tr(W_l^T D_l W_l) + \lambda_2 Tr(W_l^T D W_l), \quad (2)$$

where D_l and D are diagonal matrices with each element on the diagonal, i.e., $d_{ii}^{(l)}$ and d_{ii}, are respectively defined as

$$d_{ii}^{(l)} = \frac{1}{2||w_l^i||_2} \text{ and } d_{ii} = \frac{1}{2||w^i||_2}. \quad (3)$$

w_l^i and w^i are the i-th row of W_l and W respectively.

We approximate the solution of Eq. (2) by solving:

$$\min_{W_l^T W_l = I} Tr(W_l^T (E_l - \gamma B_l) W_l) + \lambda_1 Tr(W_l^T D_l W_l) + \lambda_2 Tr(W_l^T D W_l), \quad (4)$$

where the weight γ of the trace difference is approximated by $\gamma = \frac{Tr(\hat{W_l}^T E_l \hat{W_l})}{Tr(\hat{W_l}^T B_l \hat{W_l})}$ (Yang et al. 2012; Jia, Nie, and Zhang 2009) and $\hat{W}_l = \arg\min_{W_l} \frac{Tr(W_l^T E_l W_l)}{Tr(W_l^T B_l W_l)}$. Denote $M = (E_l - \gamma B_l)$, we have

$$\min_{W_l^T W_l = I} Tr(W_l^T M W_l) + \lambda_1 Tr(W_l^T (D_l + \mu D) W_l), \quad (5)$$

where $\mu = \lambda_1 / \lambda_1$. Thus, the optimal W can be obtained by alternatively solving Eq. (5) for the l-th task until convergence. We summarize the solution in Algorithm 1. Once W is obtained, we sort the d features according to $||w^i||_F$ in descending order and select the top ranked ones.

Experiments

We have collected a diversity of 4 public multi-class datasets: MIML (Zhou and Zhang 2007), USPS (Hull 1994), Protein (Wang 2002), and SensIT Vehicle (Duarte and Hen Hu 2004). For each dataset, we separate the multiple classes into two tasks to evaluate the performance of multi-task feature selection. We compare our method with the following feature selection algorithms: (1) Full Features which adopts all the features for classification. (2) Fisher Score (Duda, Hart, and Stork 2001). (3) SBMLR (Cawley, Talbot, and Girolami 2007) which is a sparse feature selection. (4) Multi-class $\ell_{2,1}$-norm Support Vector Machine (SVM-21) (Cai et al. 2011). (5) $\ell_{2,1}$-norm Least Square Regression (LSR-21) (Nie et al. 2010). (6) FSSI (Yang et al. 2013) which is a multi-task feature selection algorithm. We tune all the parameters (if any) by a "grid-search" strategy from $\{10^{-6}, 10^{-5}, \ldots, 10^5, 10^6\}$ and report the best results. We set the number of labeled data per class as 5 and randomly sample these labeled data to form the training sets. For each dataset, we repeat the sampling for 10 times and report the average results. Multi-class SVM is performed on the selected features to evaluate the feature selection performance.

Algorithm 1 Discriminative Multi-task Feature Selection

Input: Input data $(X_l, Y_l)|_{l=1}^t$ of t tasks. Parameters λ_1, λ_2.
Output: Matrix $W \in \mathbb{R}^{d \times c}$

```
Set r = 0 and initialize W_1|_{l=1}^t randomly;
W^(0) = [W_1, ..., W_t];
repeat
    l = 1;
    repeat
        U_l = M + λ_1(D_l + μD);
        W_l^(r) = [u_1, ..., u_{c_l}] were u_1, ..., u_{c_l} are the
        eigenvectors of U_l corresponding to the first c_l
        smallest eigenvalues;
        Update D_l^(r) using Eq. (3);
        l = l + 1;
    until l > t
    Update D^(r) using Eq. (3);
    W^(r+1) = [W_1, ..., W_t];
    r = r + 1;
until Convergence
Return W.
```

The comparison results are reported in Table 1. From the results we observe that our method obtains the better performance of classification based on the selected features. Because we utilize the discriminative information of each task, our method obtains the better results than that of the multi-task feature selection algorithm FSSI (Yang et al. 2013). The convergence curve are shown in Fig. 1. We can see that our algorithm converges within several iterations.

Acknowledgments

This paper was partially supported by the NSFC (under Grant 61202166) and Doctoral Fund of Ministry of Education of China (under Grant 20120032120042).

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Cai, X.; Nie, F.; Huang, H.; and Ding, C. 2011. Multi-class $\ell_{2,1}$-norm support vector machine. In *IEEE 11th International Conference on Data Mining*, 91–100.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Cawley, G.; Talbot, N.; and Girolami, M. 2007. Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in Neural Information Processing Systems* 19.

Duarte, M., and Hen Hu, Y. 2004. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing* 64(7):826–838.

Duda, R.; Hart, P.; and Stork, D. 2001. Pattern classification, 2nd edition. *New York, USA: John Wiley & Sons.*

Gao, C.; Wang, N.; Yu, Q.; and Zhang, Z. 2011. A feasible nonconvex relaxation approach to feature selection. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 356–361.

Han, Y.; Yang, Y.; and Zhou, X. 2013. Co-regularized ensemble for feature selection. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*.

Hull, J. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(5):550–554.

Jia, Y.; Nie, F.; and Zhang, C. 2009. Trace ratio problem revisited. *Neural Networks, IEEE Transactions on* 20(4):729–735.

Ma, Z.; Yang, Y.; Cai, Y.; Sebe, N.; and Hauptmann, A. G. 2012. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In *Proceedings of the 20th ACM international conference on Multimedia*, 469–478. ACM.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in Neural Information Processing Systems* 23:1813–1821.

Obozinski, G.; Taskar, B.; and Jordan, M. I. 2006. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Wang, J. 2002. *Application of support vector machines in bioinformatics*. Ph.D. Dissertation, National Taiwan University.

Yang, Y.; Shen, H.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, 1589–1594.

Yang, Y.; Nie, F.; Xu, D.; Luo, J.; Zhuang, Y.; and Pan, Y. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4):723–742.

Yang, Y.; Ma, Z.; Hauptmann, A.; and Sebe, N. 2013. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia* 15(3):661–669.

Zhao, Z., and Liu, H. 2007. Semi-supervised feature selection via spectral analysis. In *Proceedings of the 7th SIAM International Conference on Data Mining*, 1151–1158.

Zhou, Z., and Zhang, M. 2007. Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems* 19:1609–1616.