Effective Learning with Joint Discriminative and Representative Feature Selection

Shupeng Wang[1], Xiao-Yu Zhang[1*], Xianglei Dang[2*], Binbin Li[1*], and Haiping Wang[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
[*] *Corresponding Authors*
zhangxiaoyu@iie.ac.cn

Abstract. Feature selection plays an important role in various machine learning tasks such as classification. In this paper, we focus on both discriminative and representative abilities of the features, and propose a novel feature selection method with joint exploration on both labeled and unlabeled data. In particular, we implement discriminative feature selection to extract the features that can best reveal the underlying classification labels, and develop representative feature selection to obtain the features with optimal self-expressive performance. Both methods are formulated as joint $\ell_{2,1}$-norm minimization problems. An effective alternate minimization algorithm is also introduced with analytic solutions in a column-by-column manner. Extensive experiments on various classification tasks demonstrate the advantage of the proposed method over several state-of-the-art methods.

Keywords: Feature Selection, Discriminative Feature, Representative Feature, Matrix Optimization, Model Learning.

1 Introduction

In machine learning, high dimensional raw data are mathematically and computationally inconvenient to handle due to the curse of dimensionality [1]-[5]. In order to build robust learning models, feature selection is a typical and critical process, which selects a subset of relevant and informative features meanwhile removes the irrelevant and redundant ones from the input high-dimensional feature space [6][7]. Feature selection improves both effectiveness and efficiency of the learning model in that it can enhance the generalization capability and speed up the learning process [8][9]. The main challenge with feature selection is to select the smallest possible feature subset to achieve the highest possible learning performance.

Classic feature selection methods fall into various categories according to the involvement of classifiers in the selection procedure [10]-[12]. Although various feature selection methods have been proposed, the major emphases are placed on the discriminative ability of the features. That is to say, the features that achieve the highest classification performance are inclined to be selected. Since the classification labels are

DOI: 10.1007/978-3-319-93713-7_59

involved in feature selection, this type of feature selection methods can be severely biased to labeled data. As we know, in practical classification applications, the dataset consists of both labeled and unlabeled data. It is usually the case that there are much more unlabeled data than labeled ones. To leverage both labeled and unlabeled data for feature selection, semi-supervised feature selection methods have been studied [13][14]. Although the information underneath unlabeled data is explored, these methods are still confined to the discriminative aspect of the features. However, a comprehensive feature selection method should further take into account the representative ability with respect to the entire dataset.

Toward this end, this paper presents a novel feature selection method to explore both discriminative and representative abilities of the features. Motivated by the previous research, discriminative feature selection is implemented on labeled data via alternate optimization. Representative feature selection is further proposed to extract the most relevant features that can best recover the entire feature set, which is formulated as a self-expressive problem in the form of $\ell_{2,1}$-norm minimization. Finally, we integrate the discriminative and representative feature selection methods into a unified process. Experimental results demonstrate that the proposed feature selection method outperforms other state-of-the-art methods in various classification tasks.

2 Discriminative Feature Selection

In this paper, we follow the conventional notations, i.e. matrices are written as boldface uppercase letters and vectors are written as boldface lowercase letters. Given a matrix $\mathbf{M} = [m_{ij}]$, its i-th row and j-th column are denoted by $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Given the labeled dataset in the form of data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and the associated label matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times c}$, where d, n and c are the numbers of features, instances (or data) and classes respectively, discriminative feature selection aims at extracting the smallest possible subset of features that can accurately reveal the underlying classification labels. This can be formulated as an optimization problem which searches for the optimal projection from the feature space to the label space with only a limited number of features involved. Denoting the projection matrix as $\mathbf{A} \in \mathbb{R}^{d \times c}$, the objective is as follows.

$$\min_{\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{A}^T \mathbf{x}_i - \mathbf{y}_i\|_2 + \alpha \|\mathbf{A}\|_{2,1} \tag{1}$$

The first term in (1) is the loss of projection, and the second term is the $\ell_{2,1}$-norm regularization with parameter α to enforce several rows of $\mathbf{A}$ to be all zero. Equation (1) can be written into the matrix format:

$$\min_{\mathbf{A}} \mathcal{L}_{\mathrm{D}}(\mathbf{A}) = \|\mathbf{X}^T \mathbf{A} - \mathbf{Y}\|_{2,1} + \alpha \|\mathbf{A}\|_{2,1} \tag{2}$$

According to the general half-quadratic framework [15] for regularized robust learning, an augmented cost function $\mathcal{J}_{\mathrm{D}}(\mathbf{A}, \mathbf{p}, \mathbf{q})$ can be introduced for the minimization of function $\mathcal{L}_{\mathrm{D}}(\mathbf{A})$ in (2).

$$\mathcal{J}_{\mathrm{D}}(\mathbf{A},\mathbf{p},\mathbf{q}) = \mathrm{Tr}[(\mathbf{X}^T\mathbf{A}-\mathbf{Y})^T\mathbf{P}(\mathbf{X}^T\mathbf{A}-\mathbf{Y})] + \alpha\mathrm{Tr}(\mathbf{A}^T\mathbf{Q}\mathbf{A}) \tag{3}$$

where $\mathbf{p}$ and $\mathbf{q}$ are auxiliary vectors, while $\mathbf{P}$ and $\mathbf{Q}$ are diagonal matrices defined as $\mathbf{P} = \mathrm{diag}(\mathbf{p})$ and $\mathbf{Q} = \mathrm{diag}(\mathbf{q})$ respectively. The operator $\mathrm{diag}(\cdot)$ places a vector on the main diagonal of a square matrix. The i-th diagonal element of $\mathbf{P}$ and $\mathbf{Q}$ are:

$$p_{ii} = \frac{1}{2\left\|(\mathbf{X}^T\mathbf{A}-\mathbf{Y})^i\right\|_2} = \frac{1}{2\|\mathbf{A}^T\mathbf{x}_i-\mathbf{y}_i\|_2} \tag{4}$$

$$q_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2} \tag{5}$$

With the vectors $\mathbf{p}$ and $\mathbf{q}$ given, we take the derivative of $\mathcal{J}_{\mathrm{D}}(\mathbf{A},\mathbf{p},\mathbf{q})$ with respect to $\mathbf{A}$, and setting the derivative to zero, and arrive at:

$$\mathbf{A}^* = (\mathbf{X}\mathbf{P}\mathbf{X}^T + \alpha\mathbf{Q})^{-1}\mathbf{X}\mathbf{P}\mathbf{Y} \tag{6}$$

Note that both $\mathbf{P}$ and $\mathbf{Q}$ are dependent on $\mathbf{A}$, and thus they are also unknown variables. Based on the half-quadratic optimization, the global optimal solution can be achieved iteratively in an alternate minimization way. In each iteration, $\mathbf{P}$ and $\mathbf{Q}$ are calculated with the current $\mathbf{A}$ according to (4) and (5) respectively, and then $\mathbf{A}$ is updated with the latest $\mathbf{P}$ and $\mathbf{Q}$ according to (6). The alternate optimization procedure is iterated until convergence. After obtaining the optimal $\mathbf{A}$, discriminative features can be selected accordingly. We first calculate the absolute value of the elements of $\mathbf{A}$ by $\mathrm{abs}(\mathbf{A})$, and then sort the rows of $\mathbf{A}$ by the sums along the row dimension of $\mathrm{abs}(\mathbf{A})$. Feature selection can subsequently be performed by retaining the k features corresponding to the top k rows of sorted $\mathbf{A}$.

3 Representative Feature Selection

As for unlabeled data, only the data matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ is available, whereas the corresponding class labels are unrevealed. In this scenario, representative, rather than discriminative, feature selection is implemented to extract a limited number of informative features that are highly relevant to the rest features. The corresponding optimization problem is a self-expressive problem, which selects a relatively small subset of features that can best recover the entire feature set with linear representation. For convenience, we denote the transpose of data matrix as the feature matrix $\mathbf{F} = \mathbf{X}^T = [\mathbf{f}_1, \dots, \mathbf{f}_d] \in \mathbb{R}^{n\times d}$, whose column vectors can be regarded as n-dimensional points in the feature space. The objective is formulated as follows to obtain the representation matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$.

$$\min_{\mathbf{B}} \textstyle\sum_{i=1}^{d} \|\mathbf{F}\mathbf{b}_i - \mathbf{f}_i\|_2 + \beta\|\mathbf{B}\|_{2,1} \tag{7}$$

Similar to (1), the first term in (7) is the loss of representation, and the second term is the $\ell_{2,1}$-norm regularization to ensure row sparsity of $\mathbf{B}$ for representative feature selection. Equation (7) is equivalent to

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_59

$$\min_{\mathbf{B}} \mathcal{L}_{\mathrm{R}}(\mathbf{B}) = \|(\mathbf{FB}-\mathbf{F})^T\|_{2,1} + \beta\|\mathbf{B}\|_{2,1} \tag{8}$$

By introducing auxiliary vectors $\mathbf{p}$ and $\mathbf{q}$, we arrive at augmented cost function:

$$\mathcal{J}_{\mathrm{R}}(\mathbf{B},\mathbf{p},\mathbf{q}) = \mathrm{Tr}[(\mathbf{FB}-\mathbf{F})\mathbf{P}(\mathbf{FB}-\mathbf{F})^T] + \beta\mathrm{Tr}(\mathbf{B}^T\mathbf{QB}) \tag{9}$$

where $\mathbf{P} = \mathrm{diag}(\mathbf{p})$ and $\mathbf{Q} = \mathrm{diag}(\mathbf{q})$, with the i-th diagonal element defined as

$$p_{ii} = \frac{1}{2\|(\mathbf{FB}-\mathbf{F})_i\|_2} = \frac{1}{2\|\mathbf{Fb}_i - \mathbf{f}_i\|_2} \tag{10}$$

$$q_{ii} = \frac{1}{2\|\mathbf{b}^i\|_2} \tag{11}$$

With the vectors $\mathbf{p}$ and $\mathbf{q}$ fixed, we set the derivative of $\mathcal{J}_{\mathrm{R}}(\mathbf{B},\mathbf{p},\mathbf{q})$ to zero. Different from DFS, the analytic solution is not directly available. However, for each i ($1 \le i \le d$), the optimal representation matrix $\mathbf{B}$ can be calculated column by column with the following close form solution:

$$\mathbf{b}_i^* = p_{ii}(p_{ii}\mathbf{F}^T\mathbf{F} + \beta\mathbf{Q})^{-1}\mathbf{F}^T\mathbf{f}_i \tag{12}$$

To achieve the global optimal solution for representative feature selection, the alternate optimization according to (10), (11) and (12) is also implemented. Similarly, representative features can be selected according to the sorted $\mathbf{B}$ with respect to the row sums of $\mathrm{abs}(\mathbf{B})$.

4 Joint Discriminative and Representative Feature Selection

As we know, the cost associated with manually labeling often renders a fully labeled dataset infeasible, whereas acquisition of unlabeled data is relatively inexpensive. As a result, the available dataset typically consists of a very limited number of labeled data and relatively much more abundant unlabeled data. In order to fully explore and exploit both labeled and unlabeled data, the feature selection algorithms discussed above should be further integrated. In this section, we introduce the Joint Discriminative and Representative Feature Selection (JDRFS) algorithm, which implements DFS and RFS successively.

We denote labeled data as $\{\mathbf{X}_L \in \mathbb{R}^{d\times n_L}, \mathbf{Y}_L \in \mathbb{R}^{n_L\times c}\}$ and unlabeled data as $\mathbf{X}_U \in \mathbb{R}^{d\times n_U}$, where n_L and n_U stand for the numbers of labeled and unlabeled data respectively. The number of features to be selected, denoted as d_{DR} ($d_{\mathrm{DR}} < d$), is specified by the user beforehand. Firstly, the DFS algorithm is carried out on labeled data $\{\mathbf{X}_L, \mathbf{Y}_L\}$. Based on the optimal projection matrix $\mathbf{A}$, the least discriminative features can be preliminarily filtered out from the original d features. In this way, the candidate features are effectively narrowed down. Secondly, the RFS algorithm is performed for further selection. Instead of merely confining to unlabeled data $\mathbf{X}_U$, the entire dataset $\mathbf{X} = [\mathbf{X}_L, \mathbf{X}_U] \in \mathbb{R}^{d\times n}$ ($n = n_L + n_U$) is involved. Assuming there are d_{D} ($d_{\mathrm{DR}} < d_{\mathrm{D}} < d$) features selected after DFS, we can trim $\mathbf{X}$ by eliminating the irrelevant features and arrive at $\mathbf{X}' \in \mathbb{R}^{d_{\mathrm{D}}\times n}$, whose rows corresponds to the retained d_{D} features. After that, RFS is implemented on $\mathbf{X}'$ to obtain optimal representation matrix $\mathbf{B} \in \mathbb{R}^{d_\mathrm{D} \times d_\mathrm{D}}$. Consequently, the most representative d_DR features are selected out of the d_D features.

5 Experiments

In order to validate the performance of the JDRFS method, several experiments on various applications are carried out. Three classic feature extraction methods (i.e. PCA, ICA, and LDA), two sparse regularized feature selection methods (i.e. RoFS [10] and CRFS [11]), and the state-of-the-art semi-supervised feature selection method (i.e. SSFS [13]) are compared. The regularized softmax regression is used as classifier.

We evaluate different feature selection methods based on the classification performance of malwares [9], images [5], and patent documents [5].

For the classification applications, we implement two sets of experiments. In the first experiment, we employ a fixed number of training data and examine the classification performance with different numbers of features selected. In the second experiment, the number of features selected is fixed and we evaluate the classification performance with gradually increasing numbers of training data.

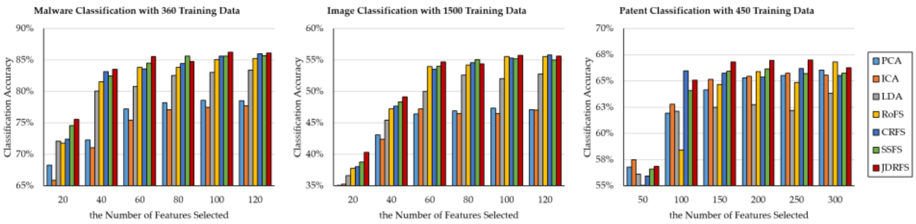

Fig. 1. The classification accuracy with fixed number of training data and different number of features selected.

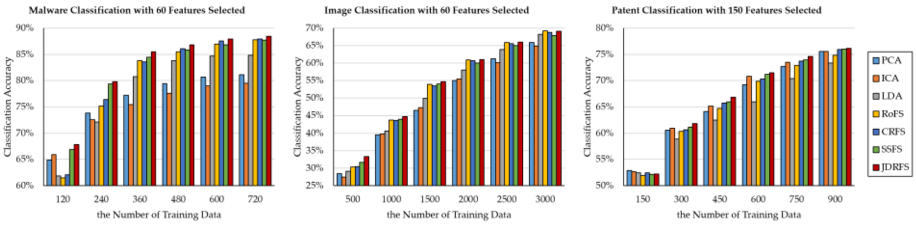

Fig. 2. The classification accuracy with fixed number of features selected and different number of training data.

Fig. 1 and Fig. 2 show the classification results corresponding to the two settings respectively. In general, JDRFS and SSFS outperform the rest methods, because they take full advantage of the information from both labeled and unlabeled data. With joint exploration on both discriminative and representative abilities of the features in an explicit way, JDRFS outperforms all the competitors and receives the highest classification accuracy. We can also see that the sparse regularized feature selection methods (CRFS, and RoFS) perform better than the classic feature extraction methods (PCA, ICA, and LDA), especially in malware and image classification. This is due to the explicit incorporation of classification labels in the feature selection objective. It also explains the higher accuracy achieved by LDA, which focuses on difference between classes of data, than PCA and ICA. As for patent classification, the advantages of CRFS and RoFS over PCA, ICA, and LDA become less significant. The most probable reason is that patent, compared with malware and image, classification is highly dependent on sophisticated domain knowledge. As a result, the classification labels offer less clue for informative feature selection. For the same reason, LDA degrades severely in patent classification.

6 Conclusion

In this paper, we have explored both labeled and unlabeled data and proposed the joint discriminative and representative feature selection method. Main contributions of this work are three-fold. Firstly, both discriminative and representative abilities of the features are taken into account in a unified process, which brings about adaptive and robust performance. Secondly, representative feature selection is proposed to extract the most relevant features that can best recover the entire feature set, which is formulated as a $\ell_{2,1}$-norm self-expressive problem. Thirdly, an alternate minimization algorithm is introduced with analytic solutions in a column-by-column manner. Extensive experiments have validated the effectiveness of the proposed feature selection method and demonstrated its advantage over other state-of-the-art methods.

7 Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant 61501457, 61602517), Open Project Program of National Laboratory of Pattern Recognition (Grant 201800018), Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (Grant 2017A05), and National Key Research and Development Program of China (Grant 2016YFB0801305).

Xiao-Yu Zhang and **Shupeng Wang** contribute equally to this paper, and are **Joint First Authors**.

References

1. Zhang, X., Xu, C., Cheng, J., Lu, H. and Ma, S., 2009. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, *11*(2), pp.272-285.
2. Liu, H. and Motoda, H., 2012. *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media.

3. Saeys, Y., Inza, I. and Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517.
4. Zhang, X., 2014. Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, *127*, pp.200-205.
5. Zhang, X.Y., Wang, S. and Yun, X., 2015. Bidirectional active learning: a two-way exploration into unlabeled and labeled data set. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(12), pp.3034-3044.
6. Liu, Y., Zhang, X., Zhu, X., Guan, Q. and Zhao, X., 2017. Listnet-based object proposals ranking. *Neurocomputing*, *267*, pp.182-194.
7. Zhang, K., Yun, X., Zhang, X.Y., Zhu, X., Li, C. and Wang, S., 2016. Weighted hierarchical geographic information description model for social relation estimation. *Neurocomputing*, *216*, pp.554-560.
8. Zhang, X.Y., 2016. Simultaneous optimization for robust correlation estimation in partially observed social network. *Neurocomputing*, *205*, pp.455-462.
9. Zhang, X.Y., Wang, S., Zhu, X., Yun, X., Wu, G. and Wang, Y., 2015. Update vs. upgrade: Modeling with indeterminate multi-class active learning. *Neurocomputing*, *162*, pp.163-170.
10. 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.
11. He, R., Tan, T., Wang, L. and Zheng, W.S., 2012. $l_{2,1}$ regularized correntropy for robust feature selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2504-2511.
12. He, R., Zheng, W.S. and Hu, B.G., 2011. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on PAMI*, *33*(8), 1561-1576.
13. Xu, Z., King, I., Lyu, M.R.T. and Jin, R., 2010. Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks*, *21*(7), 1033-1047.
14. Chang, X., Nie, F., Yang, Y. and Huang, H., 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 1171-1177.
15. He, R., Zheng, W.S., Tan, T. and Sun, Z., 2014. Half-quadratic-based iterative minimization for robust sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(2), 261-275.