LETTER

A Novel Discriminative Virtual Label Regression Method for Unsupervised Feature Selection*

Zihao SONG[†], ***Nonmember***, **Peng SONG**[†a)], ***Member***, **Chao SHENG**[†], **Wenming ZHENG**[††], **Wenjing ZHANG**[†], ***and*** **Shaokai LI**[†], ***Nonmembers***

SUMMARY Unsupervised Feature selection is an important dimensionality reduction technique to cope with high-dimensional data. It does not require prior label information, and has recently attracted much attention. However, it cannot fully utilize the discriminative information of samples, which may affect the feature selection performance. To tackle this problem, in this letter, we propose a novel discriminative virtual label regression method (DVLR) for unsupervised feature selection. In DVLR, we develop a virtual label regression function to guide the subspace learning based feature selection, which can select more discriminative features. Moreover, a linear discriminant analysis (LDA) term is used to make the model be more discriminative. To further make the model be more robust and select more representative features, we impose the $\ell_{2,1}$-norm on the regression and feature selection terms. Finally, extensive experiments are carried out on several public datasets, and the results demonstrate that our proposed DVLR achieves better performance than several state-of-the-art unsupervised feature selection methods.

key words: *feature selection, subspace learning, virtual label, linear discriminant analysis*

1. Introduction

The development of information technology produces a large number of high-dimensional data, which may contain a lot of redundant information and noises and will increase the difficulty during the data processing. Feature selection is an important technique to alleviate the influence of the curse of dimensionality. Since it can select important features and increase the speed of data processing, feature selection has attracted much attention during the past decades [1].

According to the availability of label information, feature selection can be divided into supervised feature selection, semi-supervised feature selection and unsupervised feature selection [2]. In real world, it is unrealistic to label all the data. Thus, unsupervised feature selection methods have become more popular than the other two categories. According to different selection strategies, unsupervised feature selection can be divided into the following three categories, including filter, wrapper and embedded [2]. The filtering algorithms score each feature by several fixed indexes and selects features with the highest score. However, the results are often not very satisfactory due to the simple selecting strategy. The wrapper algorithms [3] use a specific prediction and feedback learning algorithm to evaluate the feature subsets. However, with the increase of data size, the computing ability will significantly decrease and affect the performance. Embedded algorithms integrate feature selection process and learner training process [4], both of which are completed in the same optimization process. It automatically selects features in the process of learner training and can obtain more accurate results to some extent.

For unsupervised feature selection, since the label information of data is not available, it is much more challenging than the other two categories. To solve this problem, many unsupervised feature selection algorithms have been presented. For example, in [5], He et al. propose a Laplacian score (LS) algorithm for feature selection, which takes into account the local structure of data for feature selection. In [6], Cai et al. present a multi-cluster feature selection (MCFS) algorithm, which employs the spectral clustering theory to explore the manifold structure of data. In [7], Shang et al. propose a subspace learning based graph regularized feature selection (SGFS) method, which uses the graph theory to consider the local structure of feature space. In [8], Shang et al. develop a local discriminative based sparse subspace learning algorithm (LDSSL) for feature selection, which combines subspace learning with local discrimination model, and employs the ℓ_1-norm for feature selection. However, they cannot fully utilize the discriminant information of data and the virtual label information of the data.

To solve the problems mentioned above, in this letter, we propose a discriminative virtual label regression based unsupervised feature selection method. Firstly, a novel virtual label regression function is presented to guide feature selection. Then, the linear discriminant analysis (LDA) term is used to make the selected features be discriminative. Finally, we integrate the virtual label regression and LDA into the framework of subspace learning based feature selection.

Manuscript received August 4, 2021.
Manuscript revised September 29, 2021.
Manuscript publicized October 19, 2021.

[†]The authors are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, P.R. China.

[††]The author is with the Key Laboratory of Child Development and Learning Science, Ministry of Education, and School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China.

*This work was partially supported by the National Natural Science Foundation of China under Grant 61703360, and the Fundamental Research Funds for the Central Universities under Grants 2242021k30014 and 2242021k30059.

a) E-mail: pengsong@ytu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2021EDL8067

2. Proposed Method

In this section, we give some commonly used notations and then we introduce our proposed algorithm in detail. Given a matrix $A \in R^{n\times d}$, its i-th row and j-th column are denoted as A^i and A^j, respectively. $\mathrm{Tr}(A)$ represents the trace of A, A^T represents the transpose of A, and A^{-1} represents the inverse of A. The $\ell_{2,1}$-norm of A is defined as the sum of the ℓ_2-norm of the rows of A, which is expressed as $\|A\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{n} A_{ij}^2}$.

Let $X \in R^{n\times d}$ be the high-dimensional data, through subspace learning, we can learn a low-dimensional subspace of X that can well represent the original data [7], [9]. Its formula can be expressed as follows:

$$\begin{aligned} &\min_{l} \|X - X_I H\|_2^2 \\ &\text{s.t. } \|I\| = l \end{aligned} \tag{1}$$

where $H \in R^{l\times d}$ is the coefficient matrix, which is used for matrix reconstruction, $I \in R^{l\times l}$ is the index set, and l is the number of selected features from the original features. From the perspective of matrix factorization [10], Eq. (1) can be rewritten as follows:

$$\begin{aligned} &\min_{W,H} \|X - XWH\|_2^2 \\ &\text{s.t. } W,\ H \geq 0,\ W^T W = I_l \end{aligned} \tag{2}$$

where $W \in R^{d\times l}$ is the feature selection matrix. In the above equation, we have adopted an orthogonal constraint on W, which can ensure that each row or column has at most one non-zero value. Meanwhile, we adopt the $\ell_{2,1}$-norm to constrain W, which can ensure the sparsity of W. Then, Eq. (2) can be further rewritten as

$$\begin{aligned} &\min_{W,H} \|X - XWH\|_2^2 + \beta\|W\|_{2,1} \\ &\text{s.t. } W,\ H \geq 0,\ W^T W = I_l \end{aligned} \tag{3}$$

As is known to all, for unsupervised feature selection, the label information of samples is not available. Thus, we develop a linear regression function to predict labels, which aims to find a relationship between low-dimensional selected features and virtual labels. Then, the objective function is given as

$$\begin{aligned} &\min_{F,W,G} \|F - XWG\|_{2,1} \\ &\text{s.t. } F,\ G \geq 0,\ F^T F = I_c \end{aligned} \tag{4}$$

where $F \in R^{n\times c}$ is the one-hot encoding virtual label matrix, and $G \in R^{l\times c}$ is the regression matrix, which describes the relationship between the learned subspace and the virtual labels. We determine c according to the total number of categories of each dataset, and we do not know the category of each sample. Since in real world, a negative matrix has no practical meaning. Thus we impose a orthogonal constraint on F to maintain its physical meaning.

To make the model be more discriminative, we further introduce a LDA term, which aims to find the most discriminant direction by maximizing the ratio of inter-class and intra-class scatter matrices. We use the virtual label matrix F to guide the category learning in LDA. The objective function is given as follows:

$$\max_{W} \frac{W^T S_b W}{W^T S_W W} \tag{5}$$

According to [11], Eq. (5) can be expressed as the following optimization problem:

$$\min_{W} \mathrm{Tr}(W^T(S_W - uS_b)W) \tag{6}$$

where S_w is the between-class scatter matrix, S_b is the within-class scatter matrix, and u is a balancing parameter with very small positive values.

In summary, we can first use subspace learning based feature selection in Eq. (3) to learn a low-dimensional representation subspace of high-dimensional data. Then, we can use the linear regression function in Eq. (4) to predict the virtual labels of samples. Moreover, we can use the LDA term in Eq. (6) to better explore the discriminative information of data. Based on the above analysis, combining Eqs. (3), (4) and (6), the final objective function of DVLR is written as

$$\begin{aligned} &\min_{W,H,F,G} \|X - XWH\|_{2,1} + \alpha\|F - XWG\|_{2,1} \\ &+ \lambda\mathrm{Tr}(W^T(S_W - uS_b)W) + \beta\|W\|_{2,1} \\ &\text{s.t. } W,\ H \geq 0,\ F,\ G \geq 0,\ W^T W = I_l,\ F^T F = I_c \end{aligned} \tag{7}$$

where α, λ and β are trade-off parameters. By solving the above equation, we can get the feature selection matrix $W = [w_1, w_2, w_3, \ldots\ldots w_l]$. Finally, we arrange the features in a descending order according to the values of $\|W_i\|_2$, and use the selected matrix W to construct a new data matrix $X_{new} \in R^{n\times l}$.

To solve the objective function in Eq. (7), we develop an iterative update method. We introduce four Lagrangian operators δ, ζ, θ and η, which are used to ensure that W, H, F and G be non-negative. Then we can rewrite Eq. (7) as the following Lagrangian function:

$$\begin{aligned} &\min_{W,H,F,G} \|X - XWH\|_{2,1} + \alpha\|F - XWG\|_{2,1} + \lambda\mathrm{Tr}(W^T(S_W - uS_b)W) \\ &+ \frac{\gamma_1}{2}\|W^T W - I_l\|_2^2 + \frac{\gamma_2}{2}\|F^T F - I_c\|_2^2 + \mathrm{Tr}(\delta W^T) \\ &+ \mathrm{Tr}(\theta F^T) + \mathrm{Tr}(\eta G^T) + \mathrm{Tr}(\zeta H^T) + \beta\|W\|_{2,1} \end{aligned} \tag{8}$$

To resolve the $\ell_{2,1}$-norm, we define three matrices $P \in R^{n\times d}$, $U \in R^{d\times d}$ and $Q \in R^{n\times n}$, where $P_{ii} = \frac{1}{2\max(\|e_i\|_2,\epsilon)}$, $U_{jj} = \frac{1}{2\max(\|w_j\|_2,\epsilon)}$ and $Q_{ii} = \frac{1}{2\max(\|r_i\|_2,\epsilon)}$, in which e_i, w_i and r_i are the i-th rows of M=$X - XWH$, W and V=$F - XWG$, respectively. Thus we can convert Eq. (8) into the following form:

$$\begin{aligned} &\mathcal{L}(W,H,F,G) = \mathrm{Tr}(M^T PM) + \alpha\mathrm{Tr}((F - Z)^T Q(F - Z)) \\ &+ \frac{\gamma_1}{2}\mathrm{Tr}((W^T W - I_l)^T(W^T W - I_l)) \end{aligned}$$

Table 1 Clustering accuracy of different methods on different datasets (ACC±std%).

Methods	Baseline	LS	MCFS	SGFS	LDSSL	DVLR
ORL	52.74(±3.54)	45.00(±2.01)	52.42(±1.99)	55.50(±2.30)	55.75(±2.04)	**56.00(±2.88)**
COIL20	65.08(±3.02)	57.15(±2.50)	69.86(±2.79)	60.25(±2.38)	65.90(±2.75)	**70.56(±3.19)**
LUNGDIS	89.04(±7.52)	72.60(±3.70)	87.67(±)6.80	83.56(±6.43)	83.56(±6.90)	**89.81(±5.07)**
Isolet	62.11(±1.94)	58.92(±2.77)	57.50 (±1.72)	52.88(±1.89)	61.84(±2.00)	**64.49(±2.90)**
LUNG	70.27(±8.56)	61.21(±2.58)	65.52(±1.50)	59.11(±1.63)	65.02(±2.36)	**71.92(±2.24)**
JAFFE	79.75(±4.69)	92.47(±7.32)	88.47(±5.84)	86.38(±4.68)	90.14(±4.00)	**93.90(±4.01)**

Table 2 Normalized mutual information of different methods on different datasets (NMI±std%).

Methods	Baseline	LS	MCFS	SGFS	LDSSL	DVLR
ORL	72.92(±2.27)	65.29(±1.28)	73.66(±0.79)	72.49(±1.17)	72.95(±1.19)	**76.58(±1.16)**
COIL20	**79.01(±1.30)**	65.23(±1.00)	73.44(±1.28)	68.28(±1.00)	74.60(±1.80)	75.14(±1.19)
LUNGDIS	**83.08(±5.28)**	69.21(±3.76)	80.05(±)0.51	76.98(±4.32)	74.32(±4.50)	81.65(±4.37)
Isolet	76.36(±1.80)	69.67(±0.94)	69.78 (±0.75)	66.49(±1.17)	74.12(±0.89)	**76.76(±1.10)**
LUNG	54.66(±2.16)	47.42(±1.07)	50.90(±0.12)	39.33(±0.77)	46.16(±0.39)	**54.69(±2.24)**
JAFFE	83.48(±3.17)	91.88(±3.96)	87.31(±4.40)	86.32(±2.85)	87.68(±1.96)	**92.81(±3.40)**

$$+\frac{\gamma_2}{2}\mathrm{Tr}((F^TF-I_c)^T(F^TF-I_c))$$
$$+\mathrm{Tr}(\eta G^T)+\mathrm{Tr}(\delta W^T)+\mathrm{Tr}(\zeta H^T)+\mathrm{Tr}(\theta F^T)+\beta\mathrm{Tr}(W^TUW)$$
$$+\lambda\mathrm{Tr}(W^T(S_W-uS_b)W) \quad (9)$$

where $M = X - XWH$ and $Z = XWG$.
Then the optimization procedures are given as follows.

Step 1. Update W by fixing F, G and H, we can get the following equation by calculating the partial derivative of $\mathcal{L}$ with respect to W:

$$\frac{\partial\mathcal{L}}{\partial W} = 2(X^TPXWHH^T - X^TPXH^T) + \lambda(S_w - uS_b)W$$
$$+ 2\beta UW + 2\alpha(X^TQXWGG^T - X^TQFG^T)$$
$$+ 2\gamma_1(WW^TW - W) + \delta \quad (10)$$

By using the Karush-Kuhn-Tucker (KKT) conditions [12], i.e., $\delta_{ij}Wij$ =0, we can obtain

$$\Big(2(X^TPXWHH^T - X^TPXH^T) + 2\alpha(X^TQXWGG^T - X^TQFG^T)$$
$$+ 2\beta UW + 2\gamma_1(WW^TW - W) + \lambda(S_w - uS_b)\Big)W_{ij} = 0 \quad (11)$$

Let $K = \alpha X^TQXWGG^T$ and $L = X^TPXWHH^T$, we can get the update rules for W as follows:

$$W_{ij} \leftarrow W_{ij}\frac{\left[2X^TPXH^T + 2\alpha X^TQFG^T + 2\gamma_1 W + \lambda uS_bW\right]_{ij}}{\left[2L + 2K + 2\beta UW + 2\gamma_1 WW^TW + \lambda S_wW\right]_{ij}} \quad (12)$$

Step 2. Update F: We update F by fixing W, H and G. Setting the partial derivative of $\mathcal{L}$ with respect to W equal zero, we can get the following equation:

$$\frac{\partial\mathcal{L}}{\partial F} = 2\alpha(QF - QXWG) + 2\gamma_2(FF^TF - F) + \theta = 0 \quad (13)$$

By using the KKT conditions, i.e., $\theta_{ij}F_{ij}$=0, we can obtain

$$\Big(\alpha(QF - QXWG) + \gamma_2(FF^TF - F)\Big)F_{ij} = 0 \quad (14)$$

Then, we can obtain the update rule of F as follows:

$$F_{ij} \leftarrow F_{ij}\frac{[\alpha QXWG + \gamma_2 F]_{ij}}{[\alpha QF + \gamma_2 FF^TF]_{ij}} \quad (15)$$

Step 3. Update H: We update H by fixing W, F and G. By calculating the partial derivative of $\mathcal{L}$ with respect to H, we can get the following equation:

$$\frac{\partial\mathcal{L}}{\partial H} = 2(W^TX^TPXWH - W^TX^TPX) + \zeta \quad (16)$$

By using the KKT conditions, i.e., ζH=0, we can obtain the following equation:

$$(W^TX^TPXWH - W^TX^TPX)H_{ij} = 0 \quad (17)$$

Then, we can get the iterative update rules for H as follows:

$$H_{ij} \leftarrow H_{ij}\frac{\left[W^TX^TPX\right]_{ij}}{[W^TX^TPXWH]_{ij}} \quad (18)$$

Step 4. Update G: We update G by fixing W, F and H. By calculating the partial derivative of $\mathcal{L}$ with respect to G, we can get

$$\frac{\partial\mathcal{L}}{\partial G} = 2(W^TX^TQXWG - W^TX^TQF) \quad (19)$$

By using the KKT conditions, i.e., ηG_{ij}=0, we can obtain the following equation:

$$\Big(\alpha(W^TX^TQXWG - W^TX^TQF)\Big)G_{ij} = 0 \quad (20)$$

Then, we can get the update rules for G as follows:

$$G_{ij} \leftarrow G_{ij} \frac{\left[W^T X^T Q F\right]_{ij}}{\left[W^T X^T Q X W G\right]_{ij}} \tag{21}$$

3. Computational Complexity Analysis

In this section, we discuss the computational cost of the proposed DVLR method. As discussed in previous sections, the algorithm mainly includes subspace learning, virtual label regression, discriminant learning and $\ell_{2,1}$-norm constraint. Suppose n is the number of samples, d is the feature dimension, c is the number of categories of samples, and l represents the dimension of the subspace. Then, the time costs of calculating W, H, F and G are $O(d^2n + d^2c)$, $O(dnl + dl^2)$, $O(cn^2)$ and $O(dnc + dc^2)$, respectively. Due to the dimension of subspace and the number of data are much smaller than the dimension of data, the total time complexity of the DVLR is about $O(T(d^2n + d^2c))$.

4. Experiments

In this section, we evaluate the proposed DVLR algorithm on six public datasets, including two face image datasets, i.e., ORL [6] and JAFFE [13], two biological datasets, i.e., LUNG [14], LUNGDIS [14], one letter speech dataset, i.e., Isolet [15], and one digital image dataset, i.e., COIL20 [16]. The number of categories for ORL, JAFFE, LUNG, LUNGDIS, ISOLET and COIL20 are 40, 10, 5, 7, 26 and 20, respectively. To prove the effectiveness of the DVLR algorithm, we compare it with several classic unsupervised feature selection algorithms. They are Baseline (all features are used), LS [5], MCFS [6], SGFS [7] and LDSSL [8].

Here we give the settings of the parameters used in DVLR and compared algorithms. For LS, MCFS, SGFS and LDSSL, the number of nearest neighbors K is set to 5, and the Gaussian scale parameter σ is set to 10. For our method, we tune the range of α, β, γ_1 and γ_2 from $\{10^{-3}, 10^{-2}, \ldots, 10^2, 10^3\}$. We choose the dimension of subspace from $\{100, 200, 300, 400, 500\}$. For all methods, the number of selected features l is turned from $\{20, 30, 40, 50, 60, 70, 80, 90, 100\}$. The maximum number of iterations is set to 30. In addition, we perform k-means clustering with the selected features 40 times. In particular, note that in our algorithm, the virtual label regression term and the LDA term are two main parts. Thus, here we mainly analyze the two regularization parameters, i.e., α and

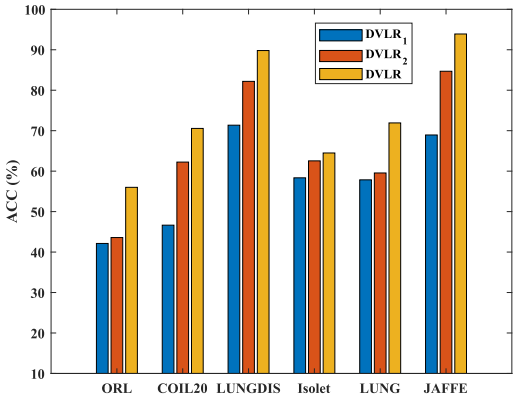
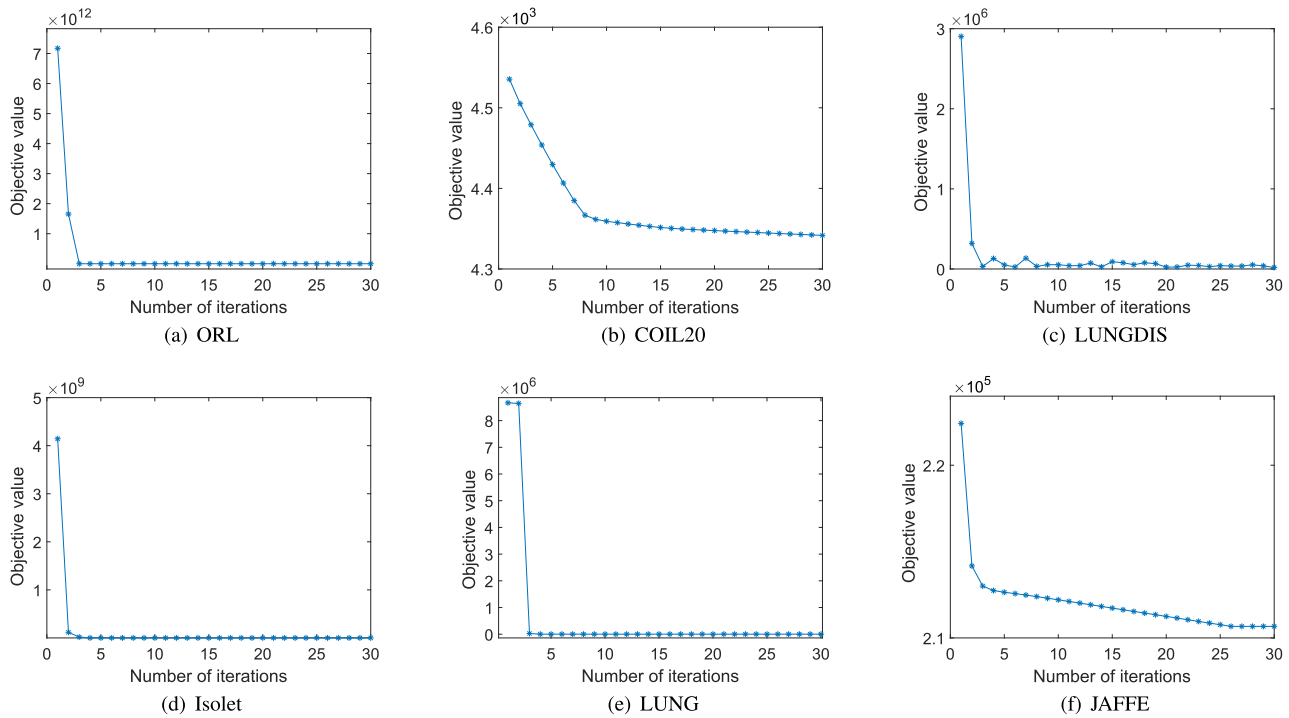


Fig. 1 ACC results of DVLR and two special cases, i.e., DVLR_1 and DVLR_2, on six datasets.



(a) ORL (b) COIL20 (c) LUNGDIS

(d) Isolet (e) LUNG (f) JAFFE

Fig. 2 Convergence curves of the proposed DVLR.

λ. For most datasets, when α and λ are set as $\{10^1, 10^2\}$ and the dimension of subspace is set as $\{300, 400\}$, the proposed method can achieve the optimal results. We choose two popular evaluation metrics, i.e., accuracy (ACC) and normalized mutual information (NMI), to evaluate the clustering results.

We give the experimental results in Tables 1 and 2, which report the ACC and NMI results, respectively. In these two tables, the best results are blacked and the second best results are underlined. From these two tables, we have the following three observations.

First, we can find that the proposed DVLR algorithm can achieve better ACC and NMI results on most datasets. In addition, in some cases, our method is better than the Baseline algorithm. These results indicate that our method can select more distinguishable and discriminative features in comparison with other methods.

Second, compared with MCFS, which uses spectral regression for feature selection, our algorithm obtains higher clustering results. The reason is that, MCFS adopts a two-step strategy. Unlike this, our algorithm integrates label regression and feature selection into a unified framework, which can achieve better results in theory.

Third, compared with SGFS and LDSSL, which also conduct on the subspace learning based feature selection framework, our method can achieve better clustering performance. The might be attributed to that, on one hand, we build a linear regression function to predict virtual labels and use the virtual labels to guild feature selection. On the other hand, we use discriminant analysis to select more discriminative features.

We further verify the effectiveness of the proposed method. By setting the regularization parameters of the virtual label regression term and the LDA term to zero, we can get a special case of DVLR, called DVLR_1. Also, by setting the the regularization parameter of the LDA term to zero, we can get a special case of DVLR, called DVLR_2. Figure 1 gives the ACC results of different cases. From the figure, we can observe that both the virtual label regression term and the LDA term can boost the clustering performance, which proves the effectiveness of our algorithm.

In addition, we experimentally study the convergence property of the proposed method. Figure 2 gives the objective values of the proposed algorithm on six datasets. From the figure, we can find that our algorithm can converge quickly on all datasets.

5. Conclusion

In this letter, we have presented an effective unsupervised feature selection method, called discriminant virtual label regression (DVLR). Different from the previous unsupervised feature selection algorithms, in DVLR, we develop a novel linear regression function to describe the linear relationship between the feature subspace and virtual labels, which can well guide the process of feature selection and discriminant subspace learning. We further integrate subspace learning, virtual label prediction and feature selection into a unified framework. Extensive experiments on six public datasets demonstrate the effectiveness of the proposed algorithm.

References

[1] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," IEEE Trans. Neural Netw. Learn. Syst., vol.28, no.7, pp.1490–1507, 2016.
[2] R. Sheikhpour, M.A. Sarram, S. Gharaghani, and M.A.Z. Chahooki, "A survey on semi-supervised feature selection methods," Pattern Recognition, vol.64, pp.141–158, 2017.
[3] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," Applied Soft Computing, vol.62, pp.441–453, 2018.
[4] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," IEEE/CAA Journal of Automatica Sinica, vol.6, no.3, pp.703–715, 2019.
[5] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," Advances in Neural Information Processing Systems, vol.18, pp.507–514, 2005.
[6] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp.333–342, 2010.
[7] R. Shang, W. Wang, R. Stolkin, and L. Jiao, "Subspace learning-based graph regularized feature selection," Knowledge-Based Systems, vol.112, pp.152–165, 2016.
[8] R. Shang, Y. Meng, W. Wang, F. Shang, and L. Jiao, "Local discriminative based sparse subspace learning for feature selection," Pattern Recognition, vol.92, pp.219–230, 2019.
[9] F.D. Mandanas and C.L. Kotropoulos, "Subspace learning and feature selection via orthogonal mapping," IEEE Trans. Signal Process., vol.68, pp.1034–1047, 2020.
[10] D.L. Sun and C. Févotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6201–6205, IEEE, 2014.
[11] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," IEEE Trans. Neural Netw., vol.17, no.1, pp.157–165, 2006.
[12] W. Xu and Y. Gong, "Document clustering by concept factorization," Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.202–209, 2004.
[13] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," IEEE Trans. Pattern Anal. Mach. Intell., vol.21, no.12, pp.1357–1362, 1999.
[14] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol.1, no.2, pp.203–209, 2002.
[15] M. Fanty and R. Cole, "Spoken letter recognition," Advances in Neural Information Processing Systems, vol.3, pp.220–226, 1990.
[16] S. Nane, S. Nayar, and H. Murase, "Columbia object image library: Coil-20," Tech. Rep., Dept. Comp. Sci., Columbia University, New York, 1996.