Adaptive Feature Selection Based on Low-Rank Representation

Ying Wang[1(✉)], Lijun Fu[1], Hongwei Zhao[2], Qiang Fu[2], Guangyao Zhai[2], and Yutong Niu[1]

[1] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
747456619@qq.com

[2] Shandong Provincial Innovation and Practice Base for Postdoctors, Shandong Baimeng Information Technology Co. Ltd., Weihai, China

Abstract. In the existing feature selection methods, the ways of construct the similarity matrix is: the first way to construct is give fixed value to two data, and the second is to calculate the distance between the two data and use it as the similarity. However, the above-mentioned method of constructing a similarity matrix is usually unreliable because the original data is often affected by noise. In the article, an adaptive feature selection method based on low-rank representation was proposed. In the method, we would dynamically construct a similarity matrix with local adaptive capabilities based on the feature projection matrix learned by the method. This construction way can reduce the influence of noise on the similarity matrix. To verify the validity of the method, we test our method on different public data sets.

Keywords: Adaptive · Feature selection · Low-rank

1 Introduction

With the advent of the information age, how to accurately classify a large amount of high-dimensional information has become an urgent problem for today's research. Feature selection is to extract features that have a greater impact on data classification from high-dimensional data. The more representative supervised feature selection methods are Relief-F [1] method proposed by K. Kira and L. A. Rendell. I. Kononenko extended the Relief-F method two years later [2]. In 2010, F. Nie et al. proposed the adaptive feature selection method [3]. In the same year, O. D. Richard et al. proposed a method for scoring feature relevance [4]. In 2012, S. Xiang et al. proposed the discriminative least squares regression method [5], which increased the discriminativeness of features by increasing the distance between different classes. In recent years, the representative semi-supervised feature selection methods include the method via spline regression method [6] proposed by Y. Han et al. and the method via rescaled linear regression method proposed by X. Chen et al. [7]. In the unsupervised feature selection method,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Shi et al. (Eds.): 6GN 2021, LNICST 439, pp. 345–353, 2022.
https://doi.org/10.1007/978-3-031-04245-4_30

because there is no label information available, the feature correlation is obtained by calculating the feature similarity. The more representative methods are Laplacian Score [8], RSFS [9] and SOGFS [10].

Among the proposed feature selection methods, the construction of similarity matrix is mostly constructed once and then the similarity matrix is unchanged. This construction method is easy to ignore the class structure, which leads to inaccurate results. Most unsupervised feature selection methods construct the similarity matrix by calculating the similarity of the original data, but because the original data is usually susceptible to noise, the way would learn the wrong feature structure.

Because the existing feature selection methods have the above problems, we proposed a feature selection method based on low-rank representation. Our method used iterative learning to obtain the similarity matrix and feature projection matrix. When constructing the similarity matrix, the low-rank representation is used as a constraint to measure the similarity of features. Using this constraint condition can solve the influence of noise on the learning feature projection matrix, so that it performs better in classification and recognition tasks. To verify the validity of the method, we test our method on several public data sets, and the experimental results are good. The main results of our paper are as follows:

1. We proposed a feature selection method based on low-rank representation. The method used low-rank representation constraints as a similarity measure. Using this similarity measure can make our feature selection method more dynamic and adaptability.
2. In our objective function, we also impose non-negative constraints on the low-rank representation coefficient, to the coefficient could dynamically and adaptively.
3. In terms of solving the objective function, we use a Lagrangian multiplier algorithm to solve the proposed method.
4. We have verified the validity of the method of our method on multiple public data sets.

The structure of the article is as follows: The second part is a review of current feature selection methods. The third part proposes our objective function and solution strategy. The fourth part describes our experimental results and analyzes the experimental results. The fifth part is a summary of this article.

2 Related Work

2.1 Low Rank Representation

Because low-rank representation can eliminate the influence of noise on sample data, low-rank representation has been applied in many fields since it was proposed, such as subspace learning, image processing and so on. The low-rank representation model is as follows:

$$\begin{aligned} &\|Z\|_* + \lambda\|E\|_{2,1} \\ &s.t.\ \ X = XZ + E \end{aligned} \tag{1}$$

In the LRR, $\|\bullet\|_*$ is the nuclear norm and $\|\bullet\|_{2,1}$ is the l_{21} norm. We assume $X = [X_1, X_2, \ldots, X_k]$ is a matrix composed of raw data from k categories, Z is a self-representation matrix, E is a noise matrix, and λ is a balance parameter. LRR not only learn the subspace of the data in noisy environment, but also discover the potential structure.

2.2 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) method is widely used because it can find a more discriminative feature subspace, which has the more discriminative subspace. In this method, we let $X \in R^{d\times n}$ is a data set composed of n objects. The data set contains a total of c categories. The method can be expressed as:

$$\max_{W^T W=1} Tr(W^T S_w W)^{-1}(W^T S_b W) \tag{2}$$

Among them, S_w is the inter-class scatter matrix, and S_b is the intra-class scatter matrix.

$$S_w = \sum_{i=1}^{c} \sum_{x_i \in y_l} (x_i - \mu_l)(x_i - \mu_l)^T \tag{3}$$

$$S_b = \sum_{i=1}^{c} (\mu_l - \mu)(\mu_l - \mu)^T \tag{4}$$

In the S_w and S_b, μ is the global eigenvector, and μ_l is the lth eigenvector. In the method, Tr($\bullet$) represents the trace and W is the projection matrix. The LDA method can find a more discriminative feature subspace.

3 Our Proposed Method

In this part, we will propose an adaptive feature selection method based on low-rank representation and analyze the proposed method from the details. In order to solve our proposed method, we will adopt a numerical strategy.

3.1 The Composition of Our Approach

As previously introduced, in the traditional methods based on the LDA framework, most methods use label information to obtain the feature subspace with the more discriminative subspace. These method of constructing feature subspace can be understood as giving different sample data the same similar weight value. It is unrealistic to adopt this construction method because of the influence of some practical factors, because in real life, even data from the same category may be different due to the influence of some factors. For example, in real life, human face images may be affected by different angles and different lighting.

Therefore we proposed an adaptive feature selection method based on low-rank representation. In the method, we used the self-representation coefficient calculated based on the original data as the weight coefficient to measure the similarity of the samples. The objective function is:

$$\begin{aligned} &\min_{W,Z,E} \|Z\|_* + \lambda\|E\|_{2,1} + \tfrac{1}{2}\sum_{ij} Z_{ij}\left\|W^T(X_i - X_j)\right\|_2 + \lambda\|W\|_{2,1} \\ &s.t.\ \ X = XZ + E, Z_{ij} > 0, W^T W = I_W \end{aligned} \tag{5}$$

In the Eq. (5), $\mathrm{X} \in \mathrm{R}^{\mathrm{d}\times\mathrm{n}}$ is a data set composed of n objects from c categories, $W \in d \times m$ is a projection matrix, Z is a self-representation matrix, and E is a noise matrix. In our objective function, the influence of noise on the data can be eliminated. $\left\|W^T(X_i - X_j)\right\|_2$ represents the distance between the samples after projected into the feature subspace. As shown in the objective function, we use Z_{ij} to constrain the structural similarity of samples X_i and X_j. At the same time, we impose non-negative constraints on Z to ensure the non-negativity of feature projection distance. If the similarity between the two samples Z_{ij} is smaller, and vice versa Z_{ij} is larger. In addition, so as to reduce the influence of redundant data on feature projection, we impose orthogonal constraints $W^T W = I_W$ on the projection matrix, I_W is the unit matrix of $W \times W$.

3.2 Optimization

In this part, we adopt a numerical strategy to solve the objective function we proposed. Because the objective function minimization problem of all variables is a non-concave problem, we use the inexact ALM algorithm to get an approximate solution. Moreover, so as to better solve the minimization problem, we have introduced auxiliary variables G and J.

$$\begin{aligned} &\min_{W,Z,E,J,G} \|J\|_* + \lambda\|E\|_{2,1} + \tfrac{1}{2}\textstyle\sum_{ij} G_{ij}\left\|W^T(X_i - X_j)\right\|_2 + \lambda\|W\|_{2,1} \\ &s.t.\ \ X = XZ + E, G_{ij} > 0, W^T W = I_W, Z = J, Z = G \end{aligned} \tag{6}$$

The augmented Lagrangian form of Eq. (6) is as follows:

$$\begin{aligned} \mathcal{L}(Z, W, M, J, G) = {} & \|J\|_* + \lambda_1\|E\|_{2,1} + \tfrac{1}{2}\textstyle\sum_{ij} G_{ij}\left\|W^T(X_i - X_j)\right\|_2 \\ & + \ \lambda_2\|W\|_{2,1} + Tr(Y_1(X - XZ - E)) + Tr(Y_2(Z - G)) \\ & + Tr(Y_3(W^T W - I_W)) + \ \ Tr(Y_4(Z - J)) + \\ & \tfrac{\mu}{2}(\|X - XZ - E\|_F^2 + \|Z - G\|_F^2 + \|Z - J\|_F^2) \\ & s.t.\ \ G_{ij} > 0 \end{aligned} \tag{7}$$

Among them, Y_1, Y_2, Y_3 are Lagrange multipliers.

In the ALM method, the minimization problem can be solved iteratively by fixing variables that have nothing to do with the solving variables. In order to solve our objective function, we first fixed G, so Eq. (6) can be transformed into:

$$\begin{aligned} &\min_{\mathrm{G}} \left\|\mathrm{Z}^{\mathrm{k}+1} - \mathrm{G} + \frac{\mathrm{Y}_2^{\mathrm{k}}}{\mu}\right\|_{\mathrm{F}}^2 + \sum_{\mathrm{ij}} \mathrm{G}_{\mathrm{ij}}\left\|\mathrm{W}^{\mathrm{T}}(\mathrm{x}_{\mathrm{i}} - \mathrm{x}_{\mathrm{j}})\right\|_2^2 \\ &\qquad \text{s.t. } \mathrm{G}_{\mathrm{ij}} \geq 0 \end{aligned} \tag{8}$$

In order to look more clear, Eq. (8) can be rewritten as:

$$\min_{G} \left\| Z^k + 1 - G + \frac{Y_2^k}{\mu} \right\|_F^2 + \sum_{ij} (R^k \otimes G_{ij}) \quad \text{s.t. } G_{ij} \geq 0 \tag{9}$$

Because each variable in formula (9) is non-negative and can be calculated independently, the minimization problem of formula (9) is a weighted norm minimization problem. This kind of problem can be mentioned in the literature [11] to solve.

In order to solve J, we eliminate variables unrelated to J, we can get

$$\min_{J} \frac{1}{\mu} \|J\|_* + \frac{1}{2} \left\| J - (Z^k + Y_4^k/\mu) \right\|_F^2 \tag{10}$$

Equation (10) is a rank minimization problem, which could be obtained through the singular value method in [8].

In order to solve W, we delete variables irrelevant to W, then we can get:

$$\min_{W} \frac{1}{2} \sum_{ij} G_{ij} \left\| W^T (x_i - x_j) \right\|_2^2 + \lambda \|W\|_{2,1} + \mathrm{Tr}(Y_3 (W^T W - I)) \tag{11}$$

Because $\|W\|_{2,1}$ in Eq. (11) is equivalent to $\sum_{l=1}^{d} \sqrt{W^l (W^l)^T}$, Eq. (11) can be rewritten as:

$$\min_{W} \frac{1}{2} \sum_{ij} G_{ij} \left\| W^T (x_i - x_j) \right\|_2^2 + \lambda \sum_{l=1}^{d} \sqrt{W^l (W^l)^T} + \mathrm{Tr}(Y_3 (W^T W - I)) \tag{12}$$

Calculating the partial derivative of W in Eq. (11), we get

$$\frac{\partial \mathcal{L}(W)}{\partial W} = \sum_{i,j=1}^{n} s_{ij} \frac{\partial \left\| W^T (x_i - x_j) \right\|_2^2}{\partial W} + 2\lambda Q W + W Y_3 \tag{13}$$

Among them, each element in the similarity matrix is:

$$s_{ij} = \frac{1}{G_{ij} \left\| W^T (x_i - x_j) \right\|_2^2} \tag{14}$$

In Eq. (13), Q is a diagonal matrix:

$$q_{ll} = \frac{1}{W^l (W^l)^T} \tag{15}$$

When we solve W, Eq. (13) can be transformed into:

$$\min_{W^T W} = I \left[\mathrm{Tr}(W^T X L_S X^T W) + \lambda \mathrm{Tr}(W^T Q W) \right] \tag{16}$$

Therefore, solving W can be obtained by solving the m smallest eigenvectors of $X L_S X^T + \lambda Q$.

Because, G and J are auxiliary variables, and there are constraints $Z = J, Z = G$ on it, because after solving G and J, the objective function of Z about can be obtained by fixing the remaining variables:

$$\min_{Z} Tr(Y_1(X - XZ - E)) + Tr(Y_2(Z - G)) + Tr(Y_4(Z - J)) + \frac{\mu}{2}(\|X - XZ - E\|_F^2 + \|Z - G\|_F^2 + \|Z - J\|_F^2) \tag{17}$$

Equation (17) is a second-order convex minimization problem. We only need to set the derivative of to zero to solve it.

Finally, we omit the irrelevant variables to E obtain the objective function of the noise matrix:

$$\min_{E} \frac{\lambda_1}{\mu}\|E\|_{2,1} + \frac{1}{2}\|E - (X - XZ + Y_1/\mu)\|_F^2 \tag{18}$$

By setting $\Psi = X - XZ + Y_1/\mu$, the minimization problem of the above formula could be solved through the method mentioned in [9].

Algorithm: feature subspace learning scheme

Input: training set $X, Z = J = G = 0, E \;\; Y_1 = Y_2 = Y_3 = Y_4 = 0,$

$\mu = 0.6,\ \mu_{\max} = 10^{10},\ \rho = 1.1$

Output: W,

While not convergence **do**

1. Update G^{k+1} using (9)

2. Update J^{k+1} using (10)

3. Update W^{k+1} using (16)

4. Update Z^{k+1} using (17)

5. Update E^{k+1} using (18)

6 Update the Y_1^{k+1}, $Y_2^{k+1}, Y_3^{k+1}, Y_4^{k+1}$ and μ;

end while

---

4 Experimental Results and Analysis

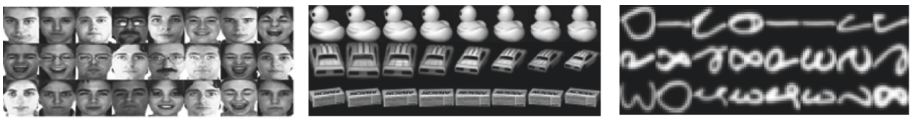
To prove the effectiveness of the method, we compared the proposed method with the existing concentrated feature subspace learning methods. The comparison methods include PCA, LDA, NPE, and LSDA.

In order to verify the effectiveness of the method, this paper uses three data sets to evaluate the adaptive feature selection model based on low-rank representation. The data sets are described as follows:

AR. AR dataset contains 3000 images of 120 targets, and each target has 26 images from different angles and illuminations. A sample picture of the data set is shown in Fig. 1(a). In the experiment, we chose 13 photos for each subject as the training set.

COIL20. The COIL20 data set contains a total of 20 objects. The camera takes pictures of each object every 5°. The data set contains a total of 1440 photos. A sample picture of the data set is shown in Fig. 1(b). When selecting training samples, the quantity of training samples for every object is 10.

USPS. The USPS dataset contains 9289 images in 10 categories. The sample image of the data set is shown in Fig. 1(c). In order to save data storage space and calculation time, the images in the data set are cropped to 16 × 16 pixels.



(a) (b) (c)

Fig. 1. (a) is sample images of AR, (b) is sample images of COIL20, (c) is sample images of USPS.

Next, we compared the ACC of the proposed method with the comparison method on the k-nearest neighbor (KNN) classifier. In our lab, the K value is set to 1. Our comparison experiment was carried out for each comparison method five times and then the standard deviation of five times was calculated. The experimental results are shown in Table 1.

According to the results, we can know that compared with the comparison, the proposed method performs better under the same conditions. Our method can find the structure of the data so as to perform label prediction on unknown label data.

Table 1. The experimental results

Methods	COIL20	AR	USPS
PCA	85.41 ± 0.32	79.12 ± 0.97	78.35 ± 1.78
LDA	84.28 ± 0.74	83.28 ± 1.46	72.53 ± 0.74
NPE	85.54 ± 1.72	81.83 ± 1.69	62.32 ± 2.21
LSDA	83.32 ± 1.63	74.27 ± 0.52	56.18 ± 2.17
Ours	92.43 ± 1.2	85.49 ± 2.21	85.32 ± 0.38

5 Conclusion

This paper proposed an adaptive feature selection method based on low-rank representation for classification. In this method, we used low-rank representation as a measure of sample similarity, and used low-rank representation as a constraint to increase the adaptability of projection space. To solve the function, we used the ALM method to solve each variable in the function. In order to verify the effectiveness of the method, we compared the performance of our proposed method with comparison on different data sets. The results show that our method performs better.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 62071157, Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011 and Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

References

1. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the 9th International Workshop Machine Learning, pp. 249–256 (1992)
2. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
3. Nie, F., Huang, H., Cai, X., Ding, C.H.: Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. In: Proceedings of Advances in Neural Information Processing System, pp. 1813–1821 (2010)
4. Richard, O.D., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, Hoboken (2010)
5. Xiang, S., Nie, F., Meng, G., Pan, C., Zhang, C.: Discriminative least squares regression for multiclass classification and feature selection. IEEE Trans. Neural Netw. Learn. Syst. **23**(11), 1738–1754 (2012)
6. Han, Y., Yang, Y., Yan, Y., Ma, Z., Sebe, N., Zhou, X.: Semisupervised feature selection via spline regression for video semantic recognition. IEEE Trans. Neural Netw. Learn. Syst. **26**(2), 252–264 (2015)
7. Chen, X., Yuan, G., Nie, F., Huang, J.Z.: Semi-supervised feature selection via rescaled linear regression. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1525–1531 (2017)

8. Candès, E.J., Li, X., Ma, Y., et al.: Robust principal component analysis? J. ACM (JACM) **58**(3), 11–49 (2011). https://doi.org/10.1145/1970392.1970395
9. Yang, J., Yin, W., Zhang, Y., et al.: A fast algorithm for edge-preserving variational multichannel image restoration. SIAM J. Imaging Sci. **2**(2), 569–592 (2009). https://doi.org/10.1137/080730421
10. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. **142**(1–2), 397–434 (2012). https://doi.org/10.1007/s10107-012-0584-1
11. Yang, J., Zhang, Y.: Alternating direction algorithms for ell_1-problems in compressive sensing. SIAM J. Sci. Comput. **33**(1), 250–278 (2011)