Robust Unsupervised Feature Selection Based on Elastic Distance Loss

Jinyan Pan¹, Youwei Xie¹, Yunlong Gao‡, Peng Gao¹, Baihua Chen² and Xiangyujie Lin¹

¹College of Information Engineering, Jimei University, Xiamen 361021, China
²Department of Automation, Xiamen University, Xiamen 361102, China
Email: gaoyl@xmu.edu.cn

Abstract. In recent years, it is a desirable way to introduce graph learning into feature selection, and it has achieved good performance in feature selection tasks. But three issues need to be concerned: (1) how to learn the similarity of samples during graph construction adaptively; (2) the performance of feature selection heavily depending on the construction of the graph; (3) most of the feature selection algorithms all use L2,1-norm for feature selection, which is not suitable for all data distributions. In this paper, we propose a robust feature selection framework to solve the above problems, which incorporated data reconstruction, graph embedding, and feature selection into a uniform framework. We use the probability of the neighbors to reconstruct the original data adaptively. Then, a novel adaptive loss function, namely, the elastic distance loss function is proposed. Finally, we compared seven state-of-art unsupervised feature selection algorithms in several benchmark datasets, and the experimental results proved the effectiveness of our proposed model.

Keywords. Unsupervised feature selection; graph learning; data reconstruction; elastic distance loss.

1. Introduction

Nowadays, unsupervised feature selection algorithms have been widely used in data mining, computer vision, and signal processing. Unsupervised feature selection aims to find the feature subsets containing a large amount of useful information. However, on account of lacking label information, it has a great challenge for unsupervised feature selection. Unsupervised Feature Selection for Multi-Cluster Data (MCFS) [1] and Laplacian Score for Feature Selection (Laplacian Score) [2] are two classical unsupervised feature selection algorithms. Laplacian Score independently evaluates the ability of each feature to retain the local manifold structure of the data without considering the correlation of each feature, gives each feature a “score”, and then selects the “high score” features. MCFS considers the importance of features in each dimension. First of all, it obtains the feature vector through spectral decomposition and then uses the \( \ell_1 - norm \) regularized regression model to perform regression analysis on the decomposed feature vector. The feature vector of the large regression coefficient is used for feature selection. Unlike MCFS and Laplacian Score, which learn the pseudo-label information of data through local manifold regularization, Robust Unsupervised Feature Selection (RUFS) [3] performs \( \ell_{2,1} - norm \) minimization in class label learning and feature learning it has better robust performance in noise samples and outlier samples. \( \ell_{2,1} - norm \) Regularized Discriminative Feature Selection for
Unsupervised Learning (UDFS) [4] uses the discriminant analysis method to find the features with the most discriminant information on all feature sets.

The algorithms mentioned above are all based on the Laplacian graph to learn the local manifold information or discriminate feature information. On the other hand, they introduce $\ell_{2,1}$-norm as a regularization term on the feature selection matrix, which could make it has row sparsity, but the graph-based feature selection algorithm has the following two shortcomings: (1) When constructing a graph to measure sample similarity, K-NN graph is used, and the choice of the size of the neighborhood will have a great impact on the graph construction. (2) Feature selection and the construction of graph structure are separated, and the quality of the final feature selection is directly affected by graph construction.

In response to the problems mentioned above, we propose a feature selection framework. In this framework, we model data reconstruction, graph embedding, and feature selection simultaneously. Based on the sample points that have a certain probability of becoming neighbors [5], we perform similarity learning based on the adaptive probability neighbors. Secondly, we consider the adaptive probabilistic neighbor to represent the data neighbor relationship while minimizing the reconstruction error. Finally, to obtain a more robust performance in the graph embedding process, we propose an elastic distance loss function. Compared with $\ell_2$-norm and $\ell_{2,1}$-norm regularization, the elastic distance loss we presented interpolated between $\ell_{2,p}$-norm and $\ell_2$-norm, so the loss of the model can be adjusted adaptively.

The rest of the paper is organized as follows: In section 2, we have a brief review of the Clustering and Projected Clustering with Adaptive Neighbors (CAN) model. In section 3, the proposed algorithm is introduced. Then, the experimental results are presented in section 4. Finally, we give a summary of the full text in section 5.

2. The CAN Model
Reference [5] proposed probabilistic neighbors to measure the similarity between sample data. Besides, they add low-rank constraints to the Laplacian matrix to adaptively learn the similarity and clustering structure. The CAN model is defined as

$$\min \sum_{ij} (\|x_i - x_j\|_2^2 s_{ij} + \gamma s_{ij}^2)$$

s.t. $s_{ij} 1 = 1, 0 \leq s_{ij} \leq 1, rank(L_s) = n - c$

where $s_{ij}$ is the probabilistic nearest neighbor graph, $L_s$ is the Laplacian matrix, $n$ is the number of samples, $c$ is the number of sample categories.

3. The Proposed Model
3.1. Elastic Distance Loss Function
In the process of model learning, if we use the $\ell_2$-norm as the loss function, it will be sensitive to outliers, but it is not sensitive to a small loss. However, if we use the $\ell_1$-norm as a loss function, it is not sensitive to large values, but it is sensitive to small loss [6]. For an ideal model, most of the sample data should have a small loss [7], while only a few data should have a large loss, which can be regarded as an outlier under this model. It can be reasonably assumed that the small loss of most data obeys the Gaussian distribution, while the larger loss of some data obeys the Laplace distribution. It is worth noting that in reference [8] pointed, when $0 < p < 1$, $\ell_p$-norm is more robust and sparse than $\ell_1$-norm, then, $\ell_{2,p}$-norm is more robust than $\ell_{2,1}$-norm, when $0 < p \leq 1$. Based on the motivation of providing adaptive loss function for different data distribution, we give the elastic adaptive loss function as follows:
\[
\text{Loss}(X) = \delta^m \| X \|_2^p + (1 - \delta)^m \| X \|_p^p
\]  

(2)

where \(0 < \delta = e^{-|x|^2} < 1, m \geq 1, 0 < p < 2\). \(\delta\) is given to balance the weight between \(\ell_{2,p} - \text{norm}\) and \(\ell_2 - \text{norm}\). The loss curve of the commonly used loss function and our proposed loss function is given in figure 1. According to figure 1, we could see that when the value \(x\) is relatively small, that is \(\delta\), is a large value, the model emphasizes \(\ell_2 - \text{norm}\), it can highlight the similarity between the sample points. On the other side, when the value \(x\) becomes larger, and the value \(\delta\) is smaller, which emphasizes \(\ell_{2,p} - \text{norm}\), which could increase the robustness to outliers.

3.2. The Objective Function

We consider the adaptive probabilistic nearest neighbors in the data reconstruction. At the same time, we also combined with graph embedding framework. The objective function is as follows:

\[
\min_{Q,A} \| Q^T X - Q^T X A \|_F^2 + \lambda \| A - S \|_F + \beta \sum_{i,j} a_{ij} \| Q^T x_i - Q^T x_j \|_2^2
\]

\[
s.t. \quad Q^T Q = I, A^T 1 = 1, A \succeq 0
\]  

(3)

where \(Q\) is the linear projection matrix, \(A\) is the weight matrix. The first item is to minimize the reconstruction error. We hope that in the linear embedding process, the data can still maintain the reconstruction relationship. The regularization term \(\lambda \| A - S \|_F\) is to force the weight matrix \(A\) to learn the similarity matrix \(S\) and avoid the emergence of trivial solutions. (Such as \(A = I\)). Secondly, to keep the manifold structure of the data well in the process of data reconstruction, the third item of the objective function considers graph embedding. Furthermore, to make the learned low-dimensional linear projection matrix more robust and sparse, the elastic adaptive loss term is introduced as the regularization term. Finally, our objective function is:

\[
\min_{Q,A} \| Q^T X - Q^T X A \|_F^2 + \lambda \| A - S \|_F + \beta \sum_{i,j} a_{ij} \| Q^T x_i - Q^T x_j \|_2^2 + \gamma \| Q \|_Q
\]

\[
s.t. \quad Q^T Q = I, A^T 1 = 1, A \succeq 0
\]  

(4)

where \(\| Q \|_Q = \delta^m \| Q \|_2^2 + (1 - \delta)^m \| Q \|_p^p\).

\[\text{Figure 1.} \quad \text{Curves of } \ell_{1} - \text{norm}, \ell_{2} - \text{norm} \quad \text{and} \quad \ell_{p} - \text{norm} \ (p = 0.5) \quad \text{loss functions and our Adaptive norm loss function.}\]
3.3. Optimization

We use alternating iterative optimization method to optimize the problem (4).

**When A is fixed**, the problem (4) is reduced to solve the following problem:

\[
\min_{\lambda} \|Q^T X - Q^T XA\|_F^2 + \beta \sum_{i,j} a_{ij} \|Q^T x_i - Q^T x_j\|_2^2 + \gamma \|Q\|_F
\]

\[
s.t. \quad Q^T Q = I
\]

Set \( R = X(I - A)(I - A)^T X^T \), \( T = XL_d X^T \), and \( D = \text{diag}(\frac{1}{\|q_i\|_2^2}) \), \( i = 1,2,\cdots,d \), where \( L_v \) is a Laplacian matrix, \( L_d = D_d - \sum\frac{a_{ij} + a_{ji}}{2} \), the equation (5) could reduce to solve the following problem:

\[
\min_{\lambda} \text{Tr}(Q^T [R + 2\beta T + \gamma [\delta^m I + (1 - \delta)^m D]]Q)
\]

\[
s.t. \quad Q^T Q = I
\]

The solution of \( Q \) is the eigenvector corresponding to the first \( m \) smaller eigenvalues of \( R + 2\beta T + \gamma [\delta^m I + (1 - \delta)^m D] \).

**When Q is fixed**, the equation (4) is reduced to solve the following problem:

\[
\min_{\lambda} \|Q^T X - Q^T XA\|_F^2 + \lambda \| A - S \|_F + \beta \sum_{i,j} a_{ij} \|Q^T x_i - Q^T x_j\|_2^2
\]

\[
s.t. \quad A^T 1 = 1, A \succeq 0
\]

Then, it can be transformed into the following problem:

\[
\min_{\lambda} \text{Tr}(BAA^T) - 2\text{Tr}(CA^T)
\]

\[
s.t. \quad A^T 1 = 1, A \succeq 0
\]

where \( B = X^T QQ^T X + \lambda r I, C = X^T QQ^T X + \lambda r S - \frac{\beta}{2} E_j^{(a)}, \tau = \frac{1}{2\|S - A\|_F}, E_j^{(a)} = \|Q^T x_i - Q^T x_j\| \).

We could solve the equation (8) by Augmented Lagrange Method, then, the equation (8) can be transformed into the following problem:

\[
\min_{\lambda} \text{Tr}(BHH^T) - 2\text{Tr}(CH^T)
\]

\[
s.t. \quad A^T 1 = 1, A \succeq 0, H = A
\]

By using the Lagrange multiplier method, we have

\[
\min_{\mu,A} \text{Tr}(BHH^T) - 2\text{Tr}(CH^T) + \frac{\mu}{2} \|A - H\|_F^2 + \langle A - H, \Theta \rangle
\]

\[
s.t. \quad A^T 1 = 1, A \succeq 0, H = A
\]

where \( \Theta \) is a Lagrange multiplier.

**When fixed A**, equation (10) is reduced to the following problem:

\[
\min_{\mu} \text{Tr}(BHH^T) - 2\text{Tr}(CH^T) + \frac{\mu}{2} \|A - H + \frac{\Theta}{\mu}\|_F^2
\]

Then, taking the partial derivative of \( H \), we have

\[
H = (2B + \mu I)^{-1}(2C + \mu A + \Theta)
\]
When fixed $H$, the equation (10) is reduced to the following problem:

$$\min_{\mu} \frac{1}{2} \| A - P \|_F^2$$

$$s.t. \ A^T I = I, A \geq 0$$

(13)

where $P = H + \Theta / \mu$, for the solution of equation (13), we solve each $a_i$ independently by the following question:

$$\min_{u_i} \frac{1}{2} \| a_i - p_i \|_2^2, i = 1, 2, ..., n$$

$$s.t. \ a_i^T I = 1, a_i \geq 0$$

(14)

Updating $\Theta, \mu$ by the following equation (15):

$$\begin{cases}
\Theta = \Theta + \mu (A - H) \\
\mu = \max(\rho \mu, u_{\text{max}})
\end{cases}$$

(15)

Then, we summarized our algorithm in table 1.

**Table 1.** The proposed Algorithm 1.

**Algorithm 1.** Robust Unsupervised Feature Selection based on Elastic Distance Loss

**Input:** Data Matrix $X \in \mathbb{R}^{d \times n}$, parameter $\lambda$, $\beta$, $\gamma$ and $p$, $\delta$ in the elastic loss function.

**Initialization:** parameter $u = 10^{-2}$, $\rho = 1.1$, $u_{\text{max}} = 10^8$, $\tau = 1$

Matrix $R = T = B = C = I$

**Step1:** Calculated similarity matrix $S$ by reference [8]

**Step2:** Calculated $Q$ by (6)

**while not converge do**

1. Updating $H$ by (12)
2. Updating $A$ by (14)
3. Updating $\Theta, \mu$ by (15)

**end while**

Calculating $R$, $T$, $B$, $C$ respectively.

Until converge

**Output:** The feature selection matrix $Q$, calculated and sorted $\| q_i \|_2^2, i = 1, 2, ..., d$, in descending order.

4. Experiments

4.1. Datasets

The datasets adopted for the experiment include the face data set ORL, Umist, Imm40; the handwriting dataset Binary Alphabet; the object dataset Coil20; the gene expression data set GLIOMA, LungD. The description of these datasets is summarized in the following table 2.
Table 2. Description of the datasets.

| Datasets     | Number of samples | Number of features | Number of classes |
|--------------|-------------------|--------------------|-------------------|
| Umist        | 575               | 1024               | 20                |
| ORL          | 400               | 1024               | 40                |
| Imm40        | 240               | 1024               | 40                |
| Binary Alphabet | 1404             | 320                | 36                |
| Coil20       | 1440              | 1024               | 20                |
| GLIOMA       | 50                | 4434               | 4                 |
| LungD        | 73                | 325                | 7                 |

4.2. Experiment Settings

We compare with the following state-of-art algorithm including MCFS [1], UDFS [4], Spectral Feature Selection for Supervised and Unsupervised Learning (SPEC) [9], RUFS [3], Joint embedding learning and sparse regression: a framework for unsupervised feature selection (JELSR) [10], Generalized uncorrelated regression with adaptive graph for unsupervised feature selection (URAFS) [11], Unsupervised Feature Selection Via Data Reconstruction and Side Information (DRSI-FS) [12]. And the parameters of all comparison algorithms are tuned in the $\{10^{-3},10^{-2},10^{-1},1,10^1,10^2,10^3\}$ number of neighbors $k$ is set to 5.

4.3. Feature Selection Experiment on Face Dataset ORL

To directly reflect the effectiveness of the proposed model, we randomly select a sample from the ORL face dataset for the feature selection experiment. Only the selected feature is shown in figure 2. The number of features from left to the right is 128, 256, 384, 512, 640, 768, 896, 1024.

According to figure 2, we can get that (1) when only a few features are selected, our model and DRSI-FS can select the main features with information (such as eyes and nose). (2) when more features are selected, our model can select all the essential information in the face (including eyes, nose, mouth, and eyebrows). Although DRSI-FS can choose eyes, nose, and mouth, the selection of eyebrow information is later than our model. (3) UDFS, SPEC, and MCFS mainly select the skin on the face but do not choose the features with discriminant information.

Figure 2. Feature selection experiment with our model, DRSI-FS, UDFS, SPEC, and MCFS algorithm (from top to bottom) in ORL face datasets. The number of features from left to right is 128, 256, 384, 512, 640, 768, 896, 1024.

4.4. Clustering Accuracy Experiment

In this experiment, the Clustering Accuracy (ACC) [13] is used to measure the performance of the algorithm under different embedded dimensions. ACC is defined following as
\[ ACC = \frac{\sum_{i=1}^{n} \delta(l_i, m_i)}{n} \]  

where \( l_i \) and \( m_i \) represent the real label and prediction label of the data, respectively. \( \delta(l_i, m_i) \) is defined in reference [13]. We first use these state-of-art methods to study the projection matrix, then, perform the k-means for the top of features as 50, 80, 110, 140, 170, 200, 230, 260, 290 for the clustering task. Through the experimental results in figure 3, our model performs better than other algorithm models in most datasets. It can be found that when the feature dimension is low, the ACC of our model and other algorithm models will be improved rapidly. With the increase of feature dimension, the accuracy of most algorithm models will decrease, and some models will decrease significantly, such as MCFS. For high-dimensional datasets, the features of discriminant information are usually in the low-dimensional subspace. That's why most algorithms have higher recognition accuracy when selecting low-dimensional features. However, when the feature dimensions increase to a certain number, noise and redundant features are inevitably introduced into the feature subset, which degrades the performance of the algorithm. Although our model also could be affected by this phenomenon, in the process of iterative optimization, we can still select features with discriminant information with the improvement of dimensions through adaptive learning.

Figure 3. The clustering recognition accuracy of algorithms on different feature dimension.

5. Conclusion
In this paper, we propose a robust unsupervised feature selection framework, which is called robust unsupervised feature selection based on elastic distance loss. Compared with the unsupervised feature selection algorithm based on graph learning and \( \ell_{2,1} - norm \) regularization, our algorithm combines data reconstruction, graph embedding learning, and feature selection into a framework. In the iterative process of the algorithm, the local manifold structure and discriminant information of the data are well preserved. At the same time, a novel loss function, the elastic distance loss function is devised, which
can adaptively adjust the loss according to the data distribution. The experimental results prove the effectiveness of our proposed model.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (42076058) and the Natural Science Foundation of Fujian Province (2020J01713).

Reference
[1] Cai D, Zhang C Y and He X F 2010 Unsupervised feature selection for multi-cluster data ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp 333-342.
[2] He X F, Cai D and Niyogi P 2005 Laplacian score for feature selection Advances in Neural Information Processing Systems 507-514.
[3] Qian M J and Cheng X Z 2013 Robust unsupervised feature selection IJCAI 1621-1627.
[4] Yang Y, Shen H T and Ma Z G 2011 L2,1-Norm regularized discriminative feature selection for unsupervised learning IJCAI 1589-1594.
[5] Nie F P, Wang X Q and Huang H 2014 Clustering and projected clustering with adaptive neighbors SIGKDD Explorations 967-976.
[6] Chen B H, Gao Y L and Wu S X 2021 Soft adaptive loss based Laplacian eigenmaps Appl. Intell. https://doi.org/10.1007/s10489-021-02300-x.
[7] Huang H, Ding C and Nie F P 2013 Adaptive loss minimization for semi-supervised elastic embedding IJCAI 1565-1571.
[8] Wang L P, Chen S C and Wang Y P 2014 A unified algorithm for mixed l (2, p)-minimizations and its application in feature selection Computational Optimization and Applications 58 (2) 409-421.
[9] Zhao Z and Liu H 2007 Spectral feature selection for supervised and unsupervised learning ICML 1151-1158.
[10] Hou C P, Nie F P, Li X L, Yi D Y and Wu Y 2014 Joint embedding learning and sparse regression: A framework for unsupervised feature selection IEEE Transactions on Cybernetics 44 (6) 793-804.
[11] LI X L, Zhang H, Zhang R, Liu Y and Nie F P 2019 Generalized uncorrelated regression with adaptive graph for unsupervised feature selection IEEE Trans. Neural Netw. Learn. Syst. 30 (5) 1587-1595.
[12] Zhang R and Li X L 2020 Unsupervised feature selection via data reconstruction and side information IEEE Transactions on Image Processing 29 8097-8106.
[13] Boriratrit S 2016 Harmonic extreme learning machine for data clustering Proc. Int. Joint Conf. Comput. Sci. Softw. Eng. pp 1-5.