A Linear Discriminant Analysis model based on Sliced Inverse Regression lifting

Boxin Nie[1, *], Xiaojia Ran[2]

[1]School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Shenyang, China, 066099

[2]School of Public Health, Hainan Medical University, Haikou, China, 571199

* Corresponding Author Email: 18310632983@163.com

Abstract. Linear Discriminant Analysis is a dimension-reduction tool for high-dimensional data in the classical field of machine learning. However, for the binary classification problem in the case of high-dimensional prediction, the dimension of the dimension reduction space of the Linear Discriminant Analysis model cannot be greater than the number of categories, that is, it can only be reduced to one dimension, which will greatly lose the effective information of the original prediction variable for the response variable. Based on this, this paper proposes an improved Linear Discriminant Analysis model based on sufficient dimensionality reduction auxiliary slice. On the one hand, by using the Sliced Inverse Regression method, the proposed method can construct auxiliary slice response variables without losing any effective information about the response variables; On the other hand, this method can increase the number of reduced dimensions of the reduced dimension space to any dimension not greater than the number of slices. The analysis results of simulated data and real data show that the proposed method is more accurate and robust than the classical Linear Discriminant Analysis model.

Keywords: Data Dimensionality Reduction, Linear Discriminant Analysis, Sliced Inverse Regression.

1. Introduction

With the increasing development of data acquisition and storage technology, more characteristics of the research object can also be recorded, reflected in the data level is different from the previous high-dimensional data. Information richness and complexity have also increased, making data processing challenging. High dimensional data often contains a lot of noise, which hinders the accurate description of the data structure and is not conducive to the description of the data structure and further exploration - this phenomenon is often called a "dimensional disaster". The most direct and effective way to alleviate this disaster is to reduce the dimension of data [1], and in the process of dimension reduction, The algorithm should try to retain the effective information of the predictive variable about the response variable.

Therefore, this paper considers supervised dimensionality reduction methods, in which Linear Discriminant Analysis (LDA), as a classical dimensionality reduction method, is applied in many fields, such as taking Linear Discriminant Analysis as the optimization goal of a deep network, learning effective hash codes by end-to-end training [2], extracting features from oil samples, making it easier to study the differences of spectral features of different samples [3], LDA on some spectral data of rice images, effectively detecting and identifying early sheath blight [4], combining LDA and PCA for accurate prediction of coronary artery disease [5], or combining with machine algorithms for identifying liquor brands [6]. LDA considers class labels or categories related to data and finds projections that maximize the separation between classes by merging class information. However, the disadvantage of traditional LDA is that the dimension reduction of Linear Discriminant Analysis cannot be greater than the number of classification tasks. Therefore, in the binary classification task, the traditional method can only set the number of dimension reduction spaces to 1. In the case of high-dimensional features, this setting will cause a large amount of original effective feature information to be lost. Therefore, there have been many researchers to improve it. For example, to reduce the impact of noise in information on dimensionality reduction and obtain a better

interpretation of characteristic variables, scholars such as Jie Wen introduced orthogonal matrix and sparse matrix and proposed robust sparse Linear Discriminant Analysis (RSLDA) [7]. To extract local information, Nie F et al. Proposed adaptive local Linear Discriminant Analysis (ALLDA) [8] or T. Tony Cai et al. Extended LDA to facilitate high-dimensional Linear Discriminant Analysis [9].

Based on this, from the perspective of sufficient dimension reduction, this paper proposes an improved Linear Discriminant Analysis based on Sliced Inverse Regression auxiliary slice. The proposed method is different from some commonly improved methods to construct auxiliary response variables. Specifically, firstly, the full dimensionality reduction technology, i.e. Sliced Inverse Regression, is used to reduce the dimension of the original data. On the premise of preserving the effective information of the original data to the greatest extent, the one-dimensional projection variable is obtained, and then the projection variable is used as the auxiliary response variable, and the LDA method is used to reduce the dimension of the feature. The innovation of this method is to construct an auxiliary variable for a slice, which can increase the dimension reduction to no more than the number of slices, and the number of slices is a super parameter, which can be large enough in theory, so it solves the problem that the traditional LDA can only reduce the dimension to one dimension. In addition, by using the idea of full dimensionality reduction, the auxiliary variables obtained should theoretically retain a large degree of information about the original characteristics of the response variables, so it is appropriate to be used as intermediate response variables. In this paper, the data processed by the proposed method and the traditional LDA method are simulated and analyzed. Specifically, the simulation data experiment results show that the proposed method can effectively improve the accuracy and robustness of the original LDA model, and the actual data analysis also shows that the proposed method has higher prediction accuracy than LDA.

2. Linear Discriminant Analysis based on sufficient dimensionality reduction auxiliary slice

To obtain the auxiliary slice response variables, firstly, the concept of sufficient dimensionality reduction is introduced. Sufficient dimensionality reduction is achieved through some linear combinations of high-dimensional independent variables in the data without making assumptions about any parameter model and without losing the information contained in the conditional distribution. For the one-dimensional response variable Y and p-dimensional explanatory variable $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})^\mathrm{T}$, the following conditions are assumed:

$$\mathrm{Y} \amalg \mathrm{X} \mid \mathrm{B}^\mathrm{T}\mathrm{X} \tag{1}$$

Where B is a $\mathrm{p} \times \mathrm{k}$-dimensional matrix, and equation (1) indicates that when $\mathrm{B}^\mathrm{T}\mathrm{X}$ is given, Y and X are conditionally independent. If $\mathrm{k} \ll \mathrm{p}$, then the purpose and effect of dimension reduction will be achieved. Therefore, the goal of full dimensionality reduction is to find a $\mathrm{p} \times \mathrm{k}$-dimensional matrix B that satisfies the above equation.

Sliced Inverse Regression (SIR) is a classical sufficient dimensionality reduction method, which is a nonparametric dimensionality reduction method proposed by Professor Li Kezhao in 1991 [10]. It can achieve dimensionality reduction without assuming any parameters and missing the relationship information between variables. For the regression problem of one-dimensional response variable Y concerning the p-dimensional explanatory variable $X = (X_1, X_2, \ldots, X_p)^T$, SIR is based on the following dimension reduction model:

$$Y = f\left(\beta_1^T X, \beta_2^T X, \ldots, \beta_p^T X; \varepsilon\right) \tag{2}$$

Of which, $\beta_1, \beta_2, \ldots, \beta_p$ is an unknown p-dimensional column vector, ε Is a random variable with a mathematical expectation of zero and independent of X. f is an arbitrary unknown function defined on $\mathbb{R}^{p+1}$. If equation (2) holds, the initial p-dimensional explanatory variable X can be replaced by k-dimensional projection $f(\beta_1^T X, \beta_2^T X, \ldots, \beta_p^T X; \varepsilon)$. If k is much less than p, then the

purpose of dimension reduction can be achieved. Therefore, it is necessary to find the matrix $B = (\beta_1, \beta_2, \dots, \beta_p)^T$ that meets the model, and matrix B has the least dimension, the least linear combination can be used to contain the information of the whole regression. It can be seen from the above that such a B matrix is not unique, so we should note that the center is reduced to the subspace $S_{Y|X}$. To ensure that the inverse regression curve falls into the central reduced dimension subspace, SIR needs a linear conditional premise: for any b in the p-dimensional space, there are constants $c_0, c_1, \dots, c_p$ to make the following equation true:

$$E(bx \mid \beta_1^T x, \beta_2^T x, \dots, \beta_p^T x) = c_0 + c_1 \beta_1^T x + c_2 \beta_2^T x + \cdots + c_p \beta_p^T x \tag{3}$$

Therefore, when equation (2) and equation (3) are satisfied, the centralized inverse regression curve $E(X|Y) - E(x)$ falls into the linear subspace generated by $\beta_k \Sigma_{xx}$, where Σ_{xx} represents the covariance matrix of X and X.

In this paper, the auxiliary slice response variable is constructed by using full dimensionality reduction, that is, the original feature dimension is reduced to one dimension by Sliced Inverse Regression, and then the auxiliary response variable is sliced, and the slice label is used as a new response variable and the classical LDA Algorithm is executed. The traditional LDA binary classification method can only reduce the dimension of discrete data to one dimension, which may lose some characteristics of the initial data, and then affect its properties and subsequent analysis. The SIR-LDA, which first improves the data by Sliced Inverse Regression, can reduce the data to any dimension not greater than the number of slices, so it can be used as a generalization of LDA model. The specific process is shown in Figure 1 below:

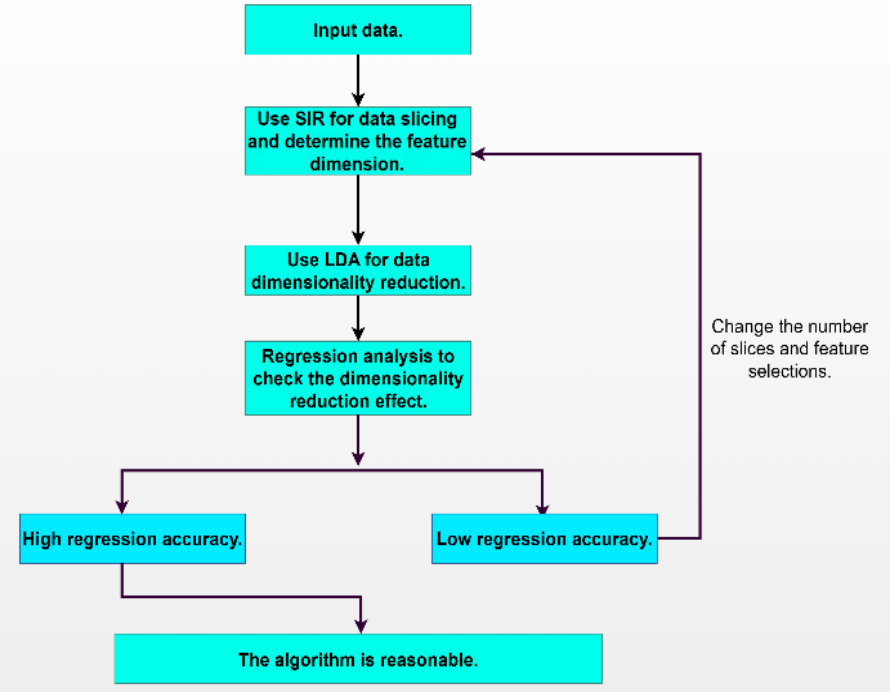


Fig. 1 SIR-LDA Algorithm flowchart

3. Data analysis

3.1. Data sources and experimental settings

According to the method mentioned above, for the continuous data X to be processed, to make it a parameter input analysis model, it is first necessary to perform a Sliced Inverse Regression operation on it. Get the discrete X'. Then LDA and a 5-fold cross test was performed. The accuracy and standard deviation of regression analysis were compared with the traditional LDA method and finally analyzed and summarized. To verify the feasibility of this method, some data are randomly generated

for the experiment. Use the *make_classification* function in the *Python* program's *sklearn* library to generate simulation data. The parameters are given in Table 1 and Table 2 below. From 1000 to 3000 groups of simulation data were randomly generated, and the simulation data generated each time included 15, 20, 25, 30, 35, and 40-dimensional explanatory variables and 1-dimensional binary response variables. Then LDA and SIR-LDA were performed, and the reduced dimension data were used for the regression prediction of the 5-fold cross test, and each group of parameters was carried out 10 times respectively. The effect and stability of regression prediction are used to measure the degree of information retention of the 2-dimensionality reduction methods.

Table 1. Parameter setting of analog data

Sample Size (n)	Count of Simulations	Select The Front n-Dimension as The Feature (fn)	Proportion of Test Data	Number of Slices	SIR-LDA Dimension
1000 1500 2000 3000	10	15, 20, 25, 30, 35, 40	50%	5	5

Table 2. Parameter setting of real data

Experiment Number	Sample Size (n)	Count of Simulations	Select The Front n-Dimension as The Feature (fn)	Proportion of Test Data	Number of Slices	SIR-LDA Dimension
1	200	10	15	0.3	5	5
2			20			
3			25			
4			30			
5			35			
6			40			

The real data comes from the ECG data set on the UCR website (https://www.cs.ucr.edu/~eamonn/time_series_data_2018). The ECG data set is binary data, which was created and provided by R. Olszewski, he formatted many data sets including this data set, and proposed a new data classification method to better describe the data [11]. The 96-dimensional data from V2 to V97 are continuous explanatory variables. V1 is a binary response variable with values of -1 and 1 respectively, which represents whether myocardial infarction has occurred. The first 5, 10, 15, 20, 25, and 30-dimensional response variables were selected as features. The following steps are the same as the analog data section. The experimental environment is *Jupiter.notebook.6.5.4 (python3)* in the *Anaconda. navigator*. The traditional LDA Algorithm and the improved SIR-LDA code of auxiliary slicing are written, and the simulated data and real data are brought into the two algorithms for 10 times cycle experiments. The test and prediction accuracy, mean and median of standard deviation are obtained. The greater the mean of accuracy, the better the accuracy, and the smaller the mean of standard deviation, the better the robustness.

3.2. Experimental result

Table 3 and Figure 2 show the accuracy and variance results of regression prediction by LDA and SIR-LDA under various parameters. Auxiliary slicing is carried out based on traditional LDA so that the data dimension after dimension reduction is no longer limited to one dimension, which improves the retention of explanatory variable information and effectively reduces the loss of information. From the simulation results, the SIR-LDA method is better than the traditional LDA method in the accuracy and standard deviation of the mean under each parameter. It can be seen from Figure 2 that

when generating 1000, 1500, 2000, and 3000-dimensional explanatory variables, all the average and most of the median results of the improved method are better than those of the traditional LDA, indicating that the improved method with Sliced Inverse Regression can improve the efficiency of the LDA Algorithm and its robustness. In general, compared with the traditional LDA, the average accuracy of SIR-LDA is improved by 9.878%, and the average standard deviation is reduced by 12.612%. From the simulation results, the binning auxiliary slicing algorithm proposed in this paper is better than the traditional LDA Algorithm in effect.

Table 3. Results of simulated data

Experiment No.	n	fn	LDA		SIR-LDA	
			Mean	Std	Mean	Std
1	1000	15	0.705	0.125	0.811	0.095
2	1000	20	0.697	0.100	0.774	0.081
3	1000	25	0.704	0.101	0.771	0.130
4	1000	30	0.763	0.090	0.831	0.066
5	1000	35	0.749	0.126	0.790	0.122
6	1000	40	0.716	0.138	0.760	0.139
7	1500	15	0.757	0.102	0.854	0.067
8	1500	20	0.720	0.069	0.816	0.058
9	1500	25	0.744	0.107	0.805	0.119
10	1500	30	0.722	0.125	0.790	0.126
11	1500	35	0.689	0.143	0.724	0.125
12	1500	40	0.678	0.109	0.728	0.106
13	2000	15	0.743	0.137	0.832	0.137
14	2000	20	0.788	0.108	0.849	0.083
15	2000	25	0.752	0.119	0.853	0.045
16	2000	30	0.801	0.080	0.863	0.070
17	2000	35	0.752	0.125	0.814	0.100
18	2000	40	0.738	0.134	0.810	0.116
19	3000	15	0.744	0.107	0.852	0.074
20	3000	20	0.708	0.088	0.858	0.076
21	3000	25	0.761	0.084	0.859	0.067
22	3000	30	0.740	0.098	0.827	0.063
23	3000	35	0.773	0.113	0.821	0.130
24	3000	40	0.795	0.134	0.804	0.138
Average:			0.739	0.111	0.812	0.097

Next, verify whether this method can process real data. LDA and SIR-LDA dimensionality reduction was separately using ECG data sets, and regression prediction was performed. The results are shown in Table 4.

Table 4. results of real data

Experiences	fn	LDA		SIR-LDA	
		Mean	Std	Mean	Std
1	5	0.698	0.051	0.778	0.036
2	10	0.715	0.058	0.817	0.029
3	15	0.768	0.040	0.840	0.024
4	20	0.740	0.054	0.812	0.052
5	25	0.750	0.040	0.787	0.060
6	30	0.780	0.053	0.835	0.041
Average:		0.715	0.049	0.811	0.040

The following is the regression test accuracy chart of dimensionality reduction data after SIR-LDA and LDA dimensionality reduction for analog data (Figures 2 to 5) and ECG data set respectively. It is found that the SIR-LDA method is superior to the traditional LDA method in the accuracy of the mean and the standard deviation of the mean under each parameter when using real data. Further, the results are visualized as Figure 6.

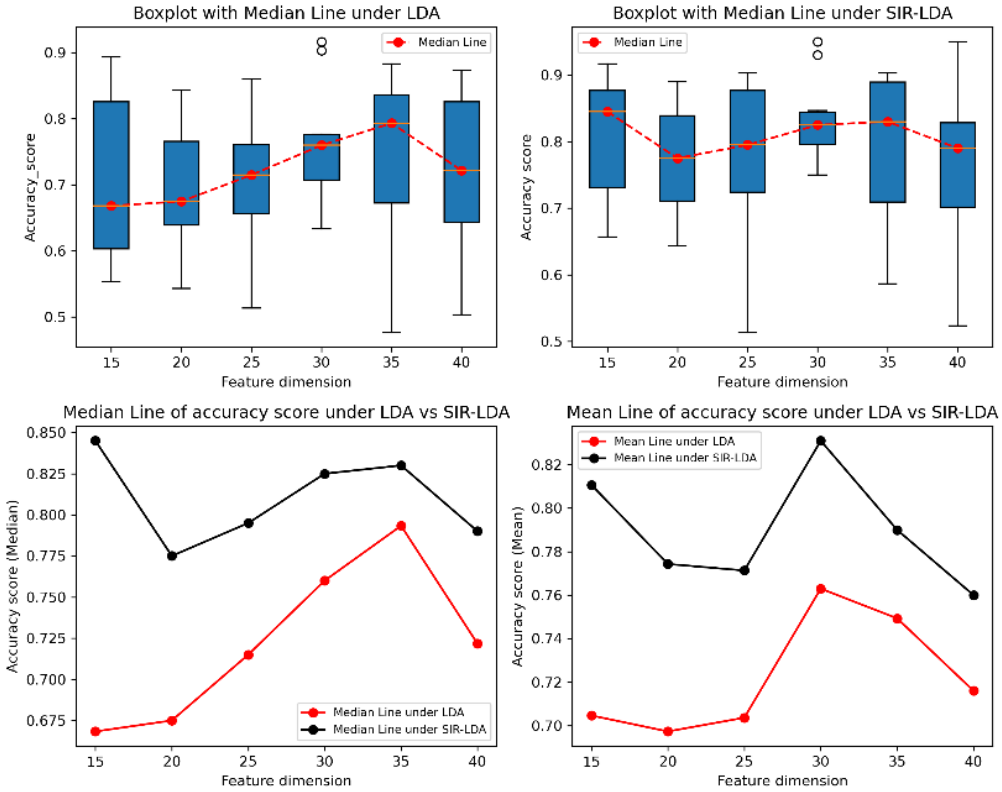


Fig. 2 Performance comparison of LDA and SIR-LDA under sample size 1000

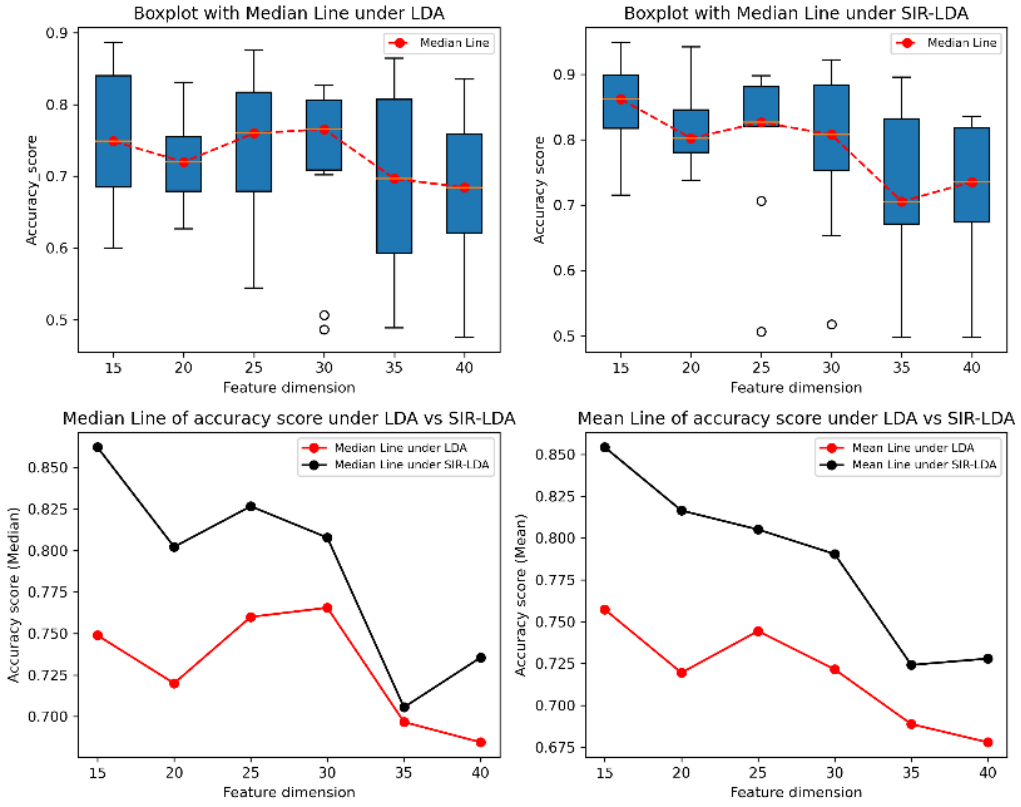


Fig. 3 Performance comparison of LDA and SIR-LDA under sample size 1500

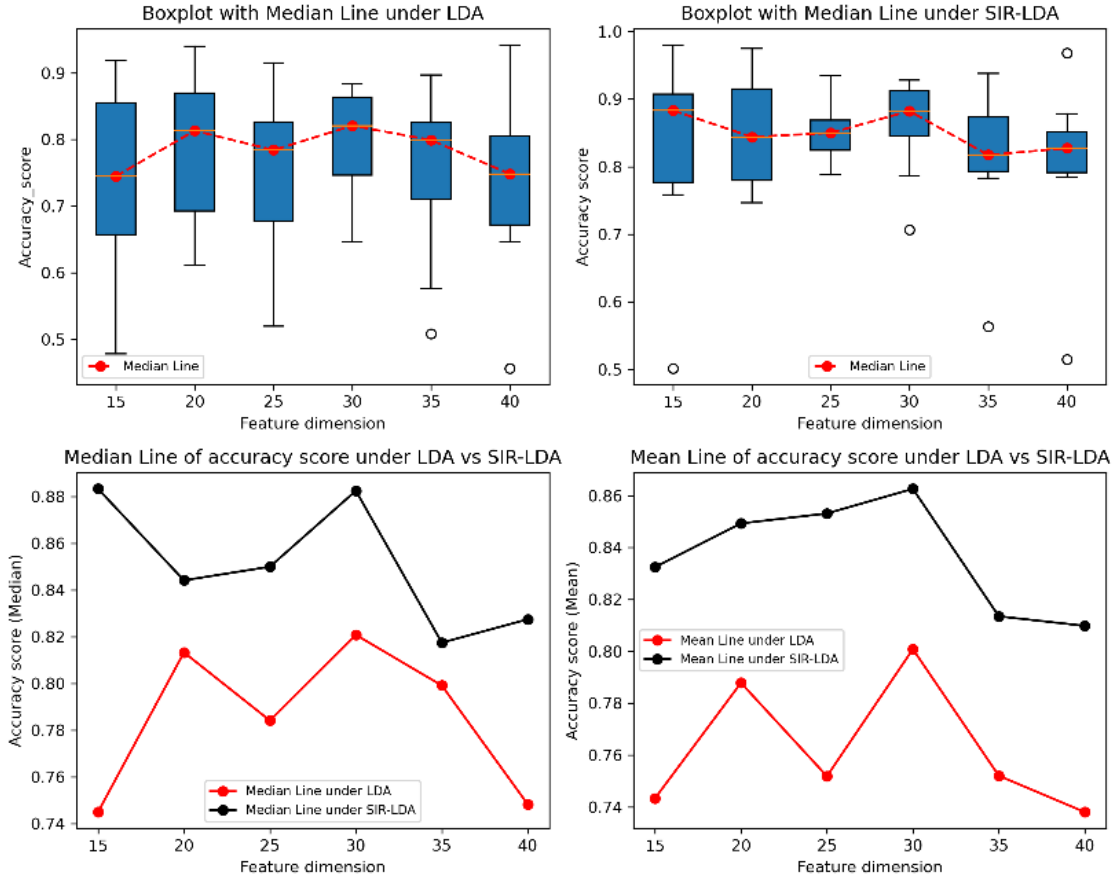


Fig. 4 Performance comparison of LDA and SIR-LDA under sample size 2000

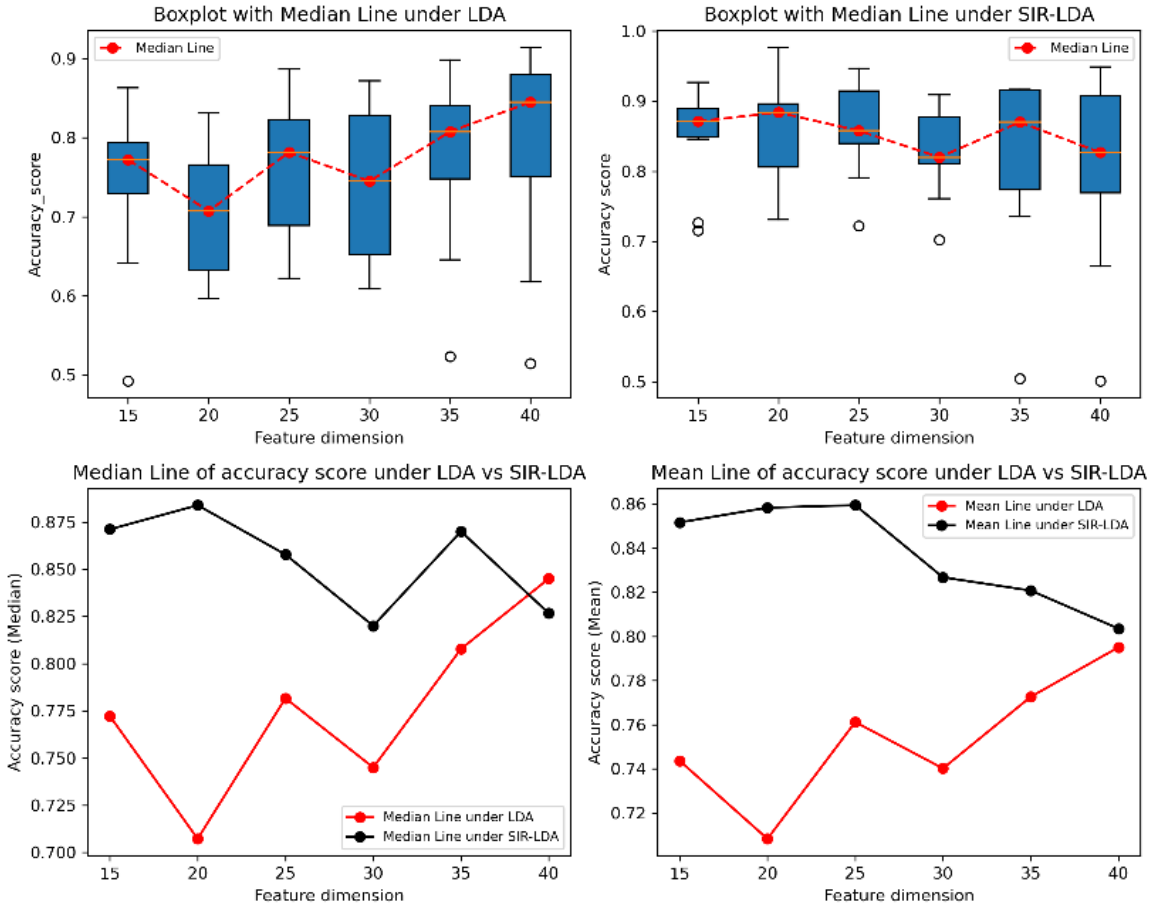


Fig. 5 Performance comparison of LDA and SIR-LDA under sample size 3000

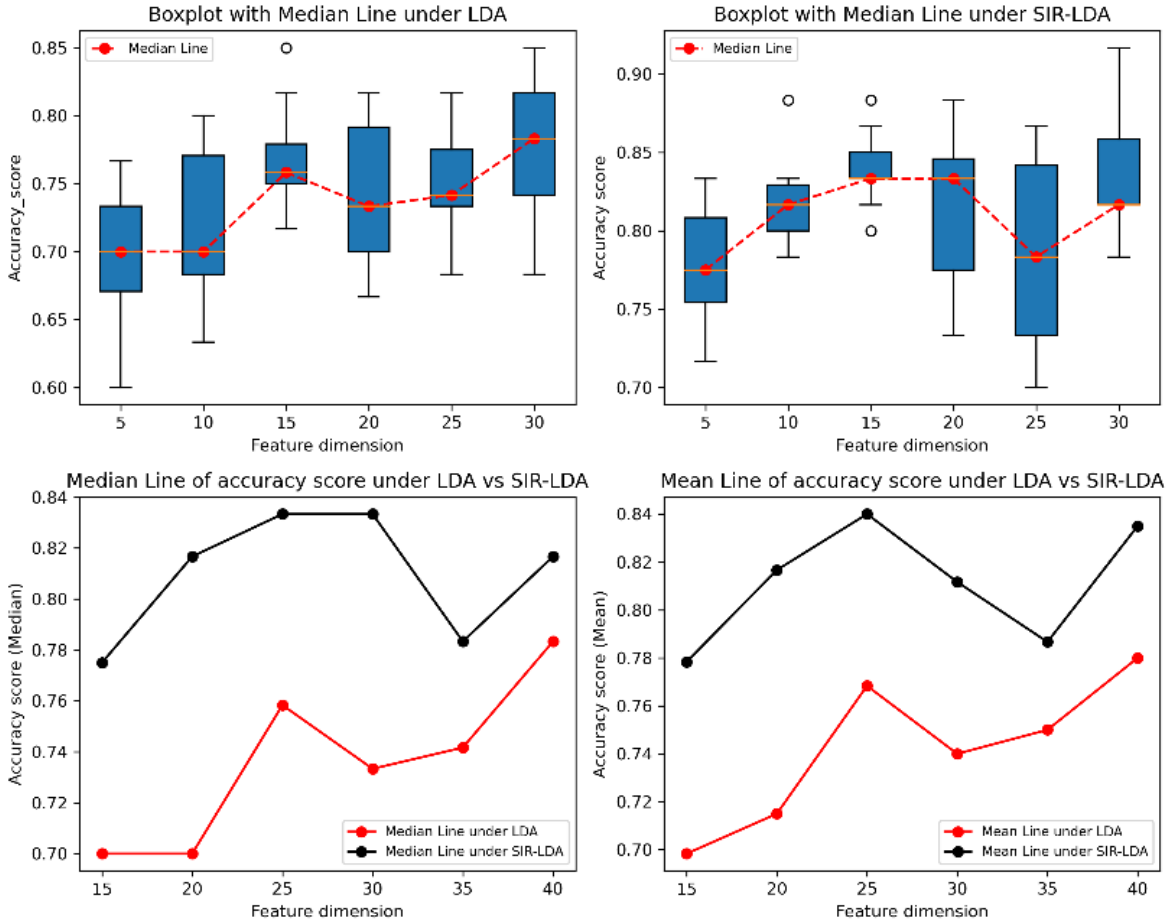


Fig. 6 Performance comparison of the LDA Algorithm and the SIR-LDA Algorithm under the ECG dataset

As can be seen from Figure 6, when performing regression analysis on the data after dimensionality reduction using SIR-LDA, the average and median accuracy rates are better than those using LDA alone. Furthermore, the box plot shows that the accuracy distribution of SIR-LDA is more concentrated, with smaller intra-group differences, indicating that the robustness has also been significantly improved. In addition, it was observed that when selecting features of different dimensions for SIR-LDA or LDA, even with the same method, there is a certain impact on the prediction accuracy. In practical applications, it is necessary to select the optimal features based on the actual problem, and the size and importance of explanatory variables (such as performing principal component analysis first).

Based on the above analysis, it can be seen that whether using simulated data or real data, using the SIR-LDA method to analyze the data after selecting the appropriate parameters can improve the accuracy and reduce the standard deviation compared to using the traditional LDA method. This indicates that the SIR-LDA method is still superior to the traditional LDA method in binary data analysis.

4. Conclusion

This article proposes an improved method for addressing the problems of traditional LDA methods. Specifically, by constructing auxiliary slice response variables through Sliced Inverse Regression, it breaks through the problem that the original LDA can only reduce the dimensionality of the response variable to 1 dimension when dealing with binary classification tasks. While constructing auxiliary slice response variables, it does not lose any effective information about the response variable from the predictive variables. From the experimental results, using SIR-LDA not only improves the overall prediction accuracy of the model but also makes the model more robust.

For future research directions, on the one hand, we need to consider using different sufficient dimensionality reduction methods as auxiliary slice response variables and discuss under what practical data scenarios it is appropriate to use. On the other hand, in the slicing part, by using equal frequency binning slicing technology, this method may encounter problems in practical data scenarios. Therefore, we need to consider using clustering slicing or other techniques that convert continuous variables into discrete variables to deal with it. Finally, the proposed method can be applied not only to binary classification tasks but also to multi-class classification tasks or continuous variable regression tasks.

References

[1] Xie Z. Research and application of data dimensionality reduction algorithm based on SIR [D] Jiangsu: Nanjing University of Posts and Telecommunications, 2018.

[2] Hu D., Nie F., Li X. Hash technology based on deep Linear Discriminant Analysis [J]. Chinese Science: Information Science, 2021, 51(02): 279-293.

[3] Kong D., Dong R., Cui Y., et al. Research on oil identification method based on three-dimensional fluorescence spectroscopy combined with two-dimensional Linear Discriminant Analysis [J]. Spectroscopy and Spectral Analysis, 2021, 41(08): 2505-2510.

[4] Zhu M., Yang H., Li Z. Early detection and recognition of rice sheath blight based on hyperspectral image and chlorophyll content [J]. Spectroscopy and Spectral Analysis, 2019, 39(06): 1898-1904.

[5] Ricciardi C., Valente A.S., Edmund K., et al. Linear Discriminant Analysis and principal component analysis to predict coronary artery disease. Health Informatics Journal. 2020;26(3):2181-2192.

[6] Huo D., Zhang M., Hou C., et al. Liquor brand identification method based on principal component analysis and discriminant analysis [J]. Transactions of the Chinese Society of Agricultural Engineering, 2011, 27(S2): 297-301.

[7] Wen, Jie, et al. "Robust sparse Linear Discriminant Analysis." IEEE Transactions on Circuits and Systems for Video Technology 29.2 (2018): 390-403. Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, and Xue long Li. 2020. Adaptive Local Linear Discriminant Analysis. ACM Trans. Knowl. Discov. Data 14, 1, Article 9 (February 2020).

[8] Nie F., Wang Z., Wang R., et al. Adaptive local Linear Discriminant Analysis[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(1): 1-19.

[9] T. Tony Cai, Lin jun Zhang, High Dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm and Missing Data, Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 81, Issue 4, September 2019, Pages 675–705.

[10] Wang Y. Application Research of face image recognition based on PCA, LDA and SVM [D]. Lanzhou Jiaotong University, 2024.

[11] Olszewski R.T. Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data[J]. Time Series Data Ph.D. Dissertation Carnagie Mellon University, 2001.