ULDA algorithm used in the study of cancer diagnosis

Yu Zhang¹,ᵃ, Yankun Li*¹,ᵇ and Kenan Huang²,ᶜ

¹School of environmental science and engineering, North China Electric Power University, Hebei 071003, China; ²The Chinese people’s liberation army 252 Hospital, Hebei 071003, China.

ᵃ1017012173@qq.com,ᵇ309267061@qq.com,ᶜknhuang369@sina.com

Keywords: ULDA; cancer diagnosis; Amino acids

Abstract. Using Uncorrelated Linear Discriminant Analysis (ULDA) algorithm, analyzing data of amino acid content in serum of cancer patients and healthy people, eventually successfully classifies the samples of cancer patients and healthy people.

1 Introduction

Cancer mortality come to first on a variety of diseases, and as the increasing of environmental pollution and the extension of human life expectancy, the incidence of cancer is increasing. Early diagnosis and timely treatment of cancer is the most effective way to improve cancer patients’ survival rate. Human serum contains rich information that can provide important clues for cancer diagnosis and treatment. But the early clinical cancer screening method based on serum has many problems such as low sensitivity, low accuracy and poor specificity, which also troubles the medical personnel and related researchers [1-3].

ULDA (Uncorrelated Linear Discriminant Analysis) was first put forward by Jin and others in the field of face recognition [4]. Now it has been successfully applied in data analysis of metabolomics, proteomics and gene expression profile [5].

In this paper, ULDA was used to find the best classification of subspace and characteristic variables, and applied to 17 kinds of amino acid content data of cancer patients and healthy people. At present, research of this aspect has not been funded though literature research. In this paper, the results showed that the model by ULDA is stable and reliable, cancer patients and healthy people were identified well. ULDA algorithm applied to the early diagnosis of cancer has a certain clinical significance and broad prospects.

2 Principle and algorithm

ULDA is a kind of feature extraction and dimension reduction algorithm based on linear discriminant analysis, aiming to maximize separate different kinds of samples, which can extract the uncorrelated linear discriminant with largest discriminant ability. At the same time, the vectors obtained are uncorrelated, which makes the information redundancy minimal [6-8].

Algorithm aims to find an optimal transformation matrix G. The data X in high dimensional space is projected into low dimensional space, as well as maximize the Fisher discriminant equation with constraint condition. Compared to the principal component vectors gotten by principal component analysis (PCA), the uncorrelated discriminant vectors (UDV) obtained from ULDA have better ability of classification. Compared to the fisher discriminant vectors gotten by Fisher discriminant analysis, uncorrelated discriminant vector are uncorrelated, which can preserve more information [5].

The traditional ULDA algorithm considers uncorrelated between column vectors on transformation matrix based on LDA, so it can reduces the data redundancy after dimension reduction. The UDV in dimension reduction space is a linear combination of the variables in an original high dimensional space, and the coefficient of combination depends on the transformation matrix G. Then the new low dimensional matrix Z can be Calculated by Z = XG.
There is the specific steps of ULDA algorithm\cite{5}.

1) Assuming that a given a data matrix \( \mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times p} \), each row of the matrix represents a sample, and each column represents a variable, also, \( n \) and \( p \) represent the number of sample and variables respectively. Assuming that the sample data belonging to the type of \( k \), The average of the whole data set is \( c_i^T \in \mathbb{R}^{1 \times p} \), superscript “T” represents the transpose of vector or matrix;

2) According to the formula \( \mathbf{H}_b = \frac{1}{\sqrt{n}} \left[ \sqrt{n}(c_1^T - c^T) \right] \) and \( \mathbf{H}_t = \frac{1}{\sqrt{n}} [X - c^T] \), \( H_b \) and \( H_t \) can be calculated;

3) Do Singular value decomposition of \( \mathbf{H}_t^T \mathbf{H}_t = \mathbf{U} \Sigma \mathbf{V}^T \);

4) \( \mathbf{U}_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \), make sure that \( u_i(i = 1, 2, \ldots, r) \) is the number of \( r \) line of matrix \( \mathbf{U} \), \( r \) is equal to the rank of \( S_t \), \( [r = \text{ranks}(S_t)] \).

5) \( \Sigma_1 = \text{diag}(\lambda_1, \ldots, \lambda_r) \), make sure that \( \lambda_i(i = 1, 2, \ldots, r) \) is the number of \( i \) element of the diagonal of matrix \( \Sigma \), as well as the number of \( i \) Nonzero eigenvalues of matrix \( H_t^T \);

6) Assuming that \( \mathbf{B} = \Sigma_1^{-1} \mathbf{U}_1^T \mathbf{H}_b^T \);

7) Do Singular value decomposition of matrix \( \mathbf{B} \), \( \mathbf{B} = \mathbf{P} \Sigma \mathbf{Q}^T \);

8) According to \( \mathbf{A} = \mathbf{U} \left[ \begin{array}{cc} \Sigma_1^{-1} P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{array} \right] \), we get matrix \( \mathbf{A} = [a_1, \ldots, a_q, a_{q+1}, \ldots, a_r] \);

9) Collect the row before \( q \) of matrix \( \mathbf{A} \) to structure transformation matrix \( \mathbf{G} \), \( \mathbf{G} = [a_1, \ldots, a_q] \), \( q \) is equal to the rank of \( S_b \), \( [q = \text{rank}(S_b)] \);

10) According to the formula \( \mathbf{Z} = \mathbf{XG} \), we can calculate a new low dimension data matrix \( \mathbf{Z} \).

3 Materials and Experimental methods

3.1 Sampling and testing

Cooperating with hospital (baoding liberation army 252 hospital) and cooperative unit (hebei agricultural university) to detect 31 cases of healthy people and 50 cases of patients’ 17 kinds of free amino acids in serum: Aspartic acid (asp D); Glutamate (glu E); Histidine (his H); Serine (ser S); Arginine (arg R); Glycine (gly G); Threonine (thr T) taurine (T), proline (pro P); Alanine (gla A); Valine (val V); methionine (met M); Cystine (cys C); Isoleucine (ile I); Leucine (leu L); Phenylalanine (phe F); Lysine (lys K);

3.2 main instruments and reagents

Agilent 1200 high performance liquid chromatograph (American Agilent company); Organization pounding machine; The thermostatic water bath pot (Beijing changfeng instrument co., LTD); PH meter (Mettle Toledo instrument co., LTD., Shanghai; LE438 pH electrode ); 17 kinds of amino acid standard reference substance.

4 The Experimental results and Data Processing

Firstly, 81 sample data were analyzed. Dividing the sample data into two part, the former 25 samples of cancer patients and former 15 samples of healthy people are used to modeling, the later 25 samples of cancer patients and later 16 samples of healthy people are used to prediction. The analysis results are shown in figure 1.
After observation, a sample deviation is founded in figure 1, judged for the singular sample. By seeking and comparison, this singular sample is the 30th sample of healthy people. Aspartic acid content of the sample value is much lower than normal, which may result from experimental error. Then delete the error data and go on.

Continue to use ULDA algorithm, analyze the rest of the 80 ample data. The former 25 samples of cancer patients and former 15 samples of healthy people are used to model. The later 25 samples of cancer patients and later 15 samples of healthy people are used to prediction. The analysis results are shown in figure 2.

From the figure 2, it can be seen, the uncorrelated discriminant vectors collected by ULDA can completely distinguish cancer patients and healthy people. The accuracy reaches 100%, which means that the results are perfect.

**Summary**

Based on the traditional ULDA algorithm, successfully used in the analysis of the amino acid content, the samples of cancer patients and healthy people are identified. It shows that ULDA has practical application significance and value to analyzing human serum common amino acids content with ULDA algorithm used for the early diagnosis of cancer.
Acknowledgements

This study is supported by the National Natural Science Foundation of China (No. 21305043); Beijing Natural Science Foundation (No. 7142102) and the Fundamental Research Funds for the Central Universities (No. 2014ZD38).

References