ULDA algorithm used in the study of cancer diagnosis

Yu Zhang[1, a], Yankun Li*[1, b] and Kenan Huang[2, c]

[1]School of environmental science and engineering, North China Electric Power University, Hebei 071003, China;

[2]The Chinese people's liberation army 252 Hospital, Hebei 071003, ChinaChina.

[a]1017012173@qq.com, [b]309267061@qq.com, [c]knhuang369@sina.com

Keywords: ULDA; cancer diagnosis; Amino acids

Abstract. Using Uncorrelated Linear Discriminant Analysis（ULDA）algorithm, analyzing data of amino acid contentin 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]. Nowit has been successfully applied in data analysis of metabolomics, proteomics and gene expression profile [5].

In this paper, ULDAwas 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 victors 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 uncorrelateddiscriminant 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 indimension 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 [5].

1) Assuming that a given a data matrix $\mathbf{X} = (x_{ij}) \in R^{n\times p}$, 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 $c_i^T \in R^{1\times p}$, superscript "T" represents the transpose of vector or matrix;
2) According to the formula $\boldsymbol{H_b} = \frac{1}{\sqrt{n}}\begin{bmatrix} \sqrt{n}(c_1^T - c^T) \\ \vdots \\ \sqrt{n_k}(c_k^T - c^T) \end{bmatrix}$ and $\boldsymbol{H_t} = \frac{1}{\sqrt{n}}[X - 1c^T]$, H_b and H_t can be calculated;
3) Do Singular value decomposition ofH_t^T,$\boldsymbol{H_t^T} = \boldsymbol{U}\sum V^T$;
4) $\boldsymbol{U_1} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_r}]$**,** make sure that$u_i(i = 1, 2, \cdots, r)$is the number of r line of matrix **U,** $r$is equal to the rank of S_t, $[r = ranks(S_t)]$.
5) $\sum_1 = diag(\lambda_1, \cdots, \lambda_r)$, make sure that $\lambda_i(i = 1, 2, \cdots, r)$ is the number of i element of the diagonal of matrix $\sum$, as well as the number of iNonzero eigenvalues of matrix H_t^T;
6) Assuming that$\mathbf{B} = \sum_1^{-1} U_1^T H_b^T$;
7) Do Singular value decomposition of matrix **B**, $\mathbf{B} = \mathrm{P}\tilde{\sum} Q^T$;
8) According to $\mathbf{A} = \mathbf{U}\begin{bmatrix} \sum_1^{-1}P & 0 \\ 0 & I \end{bmatrix}$, we get matrix $\mathbf{A} = [a_1, \cdots, a_q, a_{q+1}, \cdots, a_r]$;
9) Collect the row before q of matrix **A** to structure transformation matrix **G**, $\mathbf{G} = [a_1, \cdots, a_q]$, q is equal to the rank of S_b, $[q = rank(S_b)]$;
10) According to the formula $\mathbf{Z} = \mathbf{XG}$, we can calculate a new low dimension data matrix **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 (ger S); Arginine (arg R); Glycine (gly G); Threonine (thr T) taurine (Tau), 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 datawere 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.

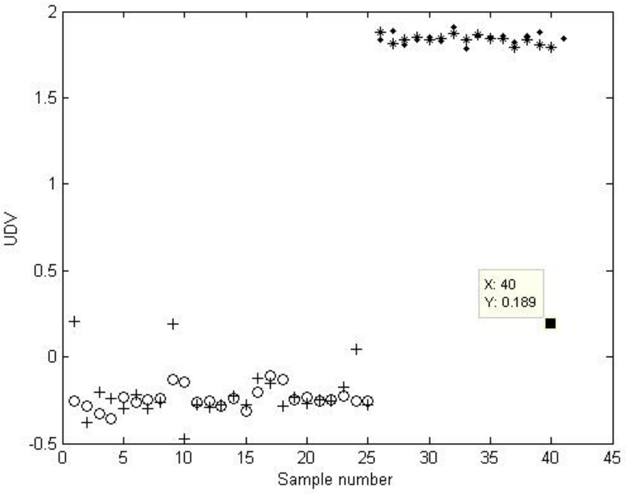

Fig. 1 Uncorrelated discriminant vectors

(“ + ”: modelingsamples of cancer patients; “ * ”:modelingsamples of healthy people; “ o ”: prediction samples of cancer patients; “ · ”:prediction samples of healthy people)

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.

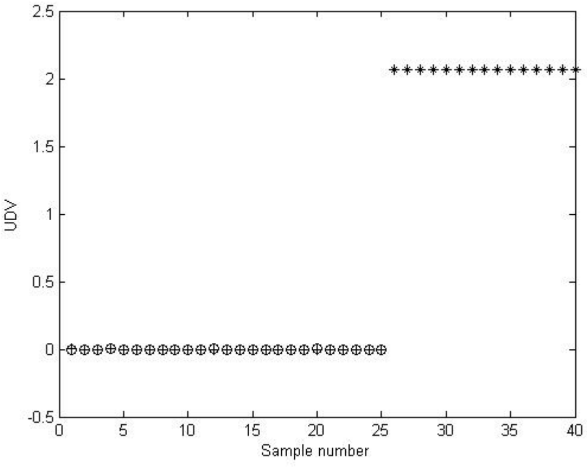

Fig.2Uncorrelated discriminant vectors

(“ + ”: model samples of cancer patients; “ * ”:model samples of healthy people; “ o ”: prediction samples of cancer patients; “ · ”:prediction samples of healthy people)

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); BeijingNatural Science Foundation (No. 7142102) and the Fundamental Research Funds for the Central Universities (No. 2014ZD38).

References

[1] Liting Yi, Jing Liu. Early detection technology research status and the latest progress of cancer. Chinese journal of medical apparatus and instruments, 2012, 36(1) : 3-51.

[2] Jing Wang. The clinical value of the combined detection of four kinds of tumor markers in serum level to lung cancer diagnosis. Practical cardio-cerebral pulmonary vascular disease, 2011(05): 753-754.

[3] Zhen Wang, Xiaodong Mei. The value of 12 tumor markers in the diagnosis of lung cancer. Guangdong medical, 2011(03): 353-356.

[4] Jin Z, Yang J-Y, Hu Z-S, et al. Face recognition based on the uncorrelated linear discriminant transformation. Pattern Recognition, 2001, 34(7):1405-1416.

[5] Yizeng Liang, QingsongXu. Complex system analysis instrument. Chemical industry press, 2012: 525-536.

[6] Qianxu Yang, Liangxiao Zhang, Longxing Wang, Hongbin Xiao. Chemometric software for multivariate data analysis based on Matlab. Chemo metrics and Intelligent Laboratory Systems, 2012,116:1-8.

[7] Delin Chu, SiongTheGoh, Y. S. Hung. Characterization of all solutions for under sampled uncorrelated linear discriminant analysis problems. Society for Industrial and Applied Mathematics, 2011, 32(3):820-844.

[8] X.H. Chen, S.C. Chen, H. Xue. Universum linear discriminant analysis. Electronics Letters, 2012, 48(22): 1407-1409.