Asia-Pacific Journal of Computer Science and Technology
亚太计算机科学与技术期刊
2019; 1(3): 28-33
http://www.sciencepublishinggroup.com/j/apjcst

Application and Effect Comparison of PCA and PLSDA in TCM Constitution

Xie Fangrong[1], Zhou Xiaoyun[1], Han Liang[2], Shi Zhongfeng[2], Chen Guanhao[2], Huang Haiquan[3], Cheng Qi[3], Chen Ziqiang[3], Hu Jinyuan[4], Song Yuhong[1], Xu Shu[5, *]

[1]First People's Hospital Affiliated to Guangzhou Medical University, Guangzhou, China
[2]Health departments,Guangdong Pharmaceutical University, Guangzhou, China
[3]Guangdong Yisheng Information Technology Co., Ltd., Guangzhou, China
[4]Guangdong Shengshijianwang Health Management Co., Ltd., Guangzhou, China
[5]Chinese Academy of Sciences University Shenzhen Hospital (Guangming), Shenzhen, China

Email address:
xiefangrong@126.com (Xie Fangrong), selflearner@126.com (Xu Shu)
[*]Corresponding author

To cite this article:
Xie Fangrong, Zhou Xiaoyun, Han Liang, Shi Zhongfeng, Chen Guanhao, Huang Haiquan, Cheng Qi, Chen Ziqiang, Hu Jinyuan, Song Yuhong, Xu Shu. Application and Effect Comparison of PCA and PLSDA in TCM Constitution. *Asia-Pacific Journal of Computer Science and Technology*. Vol. 1, No. 3, 2019, pp. 28-33.

Received: November 15, 2019; **Accepted**: February 9, 2020; **Published**: March 16, 2020

Abstract: TCM constitution refers to the individual characteristics of the structure, function, and metabolism formed by the human body on the basis of innate endowments during the growth and development and aging of the day after tomorrow. In this study, 657 people from the community were randomly selected to complete the questionnaire of TCM constitution identification and score.Through the questionnaire, the scores of 9 different constitutions (yang-deficiency, yin-deficiency, qi-deficiency, phlegm-dampness, damp-heat, blood-stasis, characteristic, qi-stagnation and placid) were obtained, and the results of constitution classification were obtained.Then PCA and PLSDA were used to compare different TCM constitutions. PCA and PLSDA are commonly used dimensionality reduction tools, and are now mostly used in statistics, mathematics, and computer science. In this study, we will use PCA and PLSDA to compare different types of TCM constitutions, use PCA and PLSDA's dimensionality reduction ideas to combine different sample data, and to achieve normalized processing of data of different dimension units, so as to construct samples with better properties. Data to better classify different TCM constitutions. This study found that the physique of the peaceful population is very different from that of other constitutional populations. It also provides a scientific and reliable proof for the application of modern data to identify the constitution of traditional Chinese medicine.

Keywords: PCA, PLSDA, TCM Constitution

PCA和PLSDA在中医体质分型中的运用及效果对比

谢方镕[1]，周晓芸[1]，韩亮[2]，石忠峰[2]，陈冠豪[2]，黄海铨[3]，程琦[3]，陈自强[3]，胡金元[4]，宋雨鸿[1]，徐舒[5*]

[1]广州医科大学附属市一人民医院，广州，中国
[2]广东药科大学，健康学院，广州，中国
[3]广东易生活信息科技有限公司，广州，中国
[4]广东盛世健王健康管理有限公司，广州，中国
[5]中国科学院大学深圳医院(光明)，深圳，中国

邮箱
xiefangrong@126.com（谢方镕），selflearner@126.com（徐舒）

摘要：中医体质是指人体以先天禀赋为基础，在后天的生长发育和衰老过程中所形成的结构、功能和代谢上的个体特殊性。PCA和 PLSDA 是常用的降维工具，现多运用于统计学、数学、计算机学等。本研究通过问卷调查的方式，随机选取社区群众657人完成《中医体质辨识问卷表》的填写，完成的得分。通过问卷完成人体9种不同体质（阳虚质、阴虚质、气虚质、痰湿质、湿热质、血淤质、特禀质、气郁质及平和质）的得分，从而得到体质分型结果。再运用PCA和PLSDA来比较不同的中医体质分型，利用 PCA 和PLSDA的降维思想结合不同的样本数据，实现对不同量纲单位数据的归一化处理，从而构建性质较好的样本数据，对不同的中医体质作出更好的分类。本研究发现平和质人群的体质和其他体质人群相差很大，同时也为现代数据的应用来辨识中医体质提供了科学性、可靠性的证明。

关键词：PCA，PLSDA，中医体质

1．中医体质的研究进展

从《黄帝内经》时代起，人们就意识到由于每个人的体质之间存在很大区别，所以诊治时需要将人按不同体质进行分类，根据各人体质类别选择不同方式的治疗。如《灵枢·行针》篇中就曾提到：“百姓之血气，各不同形，或神动而气先针行；或气与针相逢；或针已出，气独行；或数刺乃知；或发针而气逆；或数刺病益剧”，说明不同体质的人在行针之前要辨明体质，再用不同进针手法治疗[1-2]。后来，随着时代的发展，到了东汉末年，“医圣”张仲景又提出，体质存在寒、热、燥、湿、虚、实等不同，诊治前应先辨明体质。而到了现代，体质又有了很多新的分类，其中有三个流派最具代表性：一是以王琦为代表的“身心统一论”，将体质分为平和质、气虚质、阳虚质、阴虚质、痰湿质、湿热质、瘀血质、气郁质、特禀质九类；二是以何裕民为代表的“身体素质论”，将体质分为强壮型、虚弱型、偏寒型、偏热型、偏湿型、瘀迟型六型；三是以匡调元为代表的“两纲八要”分类法，将体质分为晦涩质、腻滞质、燥红质、迟冷质、倦质、正常质六种[3-5]。本文中的体质分型是按王琦教授提出的九分法来进行分型。

2．PCA与PLSDA

主成分分析（Principal Component Analysis，PCA）是一种数据压缩和特征提取的多变量统计分析技术，本研究使用 PCA 对网络的外部输入变量进行降维，通过构造不同中医体质形成的变量的一系列线性组合形成新变量，新的变量比原始数据维度更低，而且在彼此不相关的前提下反映原始数据的信息[6]。偏最小二乘法判别分析(Partial least squares discrimination analysis, PLSDA)是常用的降维工具，现已广泛运用于统计学、数学、计算机学、经济学、气象学、等多个领域的预测与鉴别[7-9]。本研究主要阐述PCA和PLSDA在中医体质分型中的运用及效果对比。

3．数据来源及处理方法

3.1．数据来源

随机选取广东省佛山市顺德区龙江镇40-70岁年龄段的社区群众657人，通过问卷调查的方式，向被调查者派发北京中医药大学王琦教授的《中医体质辨识问卷表》。

3.2．检测方法

根据被调查者填写的内容，可以得到每个人阳虚质、阴虚质、气虚质、痰湿质、湿热质、血淤质、特禀质、气郁质及平和质9种体质的得分。通过每种体质的得分，按一定标准可得到该人的体质分型结果。再将所得结果分别导入PCA及PLSDA两种模型进行降维及分类运算得出结果。

3.3．数据处理

（I）PCA法

PCA 分析，即主成分分析（Principal Component Analysis），首先是在1901年由Karl Pearson 等提出，适用于非随机变量数据，1933年，Hostelling将此方法推广到随机向量。主成分分析信息的大小通常用离差平方和或方差来衡量，它是一种通过线性变换对数据进行简化分析的技术，在数学领域经常被用于将高维数据降到低维数据，即“降维”操作。它关键步骤是将提出的所有变量中重复的或相关性大的变量删去冗余，留下或重新建立变量。通过保留低阶主成分，忽略高阶主成分，保证这些主成分之间互不相关并且最大限度地保持原有的信息的过程来实现的[10-11]。

（II）PLSDA 法

偏最小二乘法判别分析（Partial Least Squares Discriminant Analysis, PLSDA） PLSDA 是基于偏最小二乘（Partial Least Squares，PLS）回归的分类方法，分类能力很强。是一种常用于判别分析多变量的统计分析方法[12-13]。该方法结合了多元线性回归和主成分分析的优点，即使在数据相关性较小的情况下，也能得到较为准确的结果。

（III）GBDT 法

除了对模型的拟合参数进行对比外，为了进一步对比分析PCA和PLSDA两种模型在中医体质辨识效果项目中的降维可信度，我们引入了目前公认且最常用分类器：GBDT分类器。通过对比PCA+GBDT、PLSDA+GBDT和原始特征+GBDT这三种处理方法的分类效果，采用Accuracy score作为评价指标来评价PCA与PLSDA的降维效果。

GBDT, 即梯度提升决策树（Gradient Boosted Decision Trees）。该分类器的算法核心是在每一次迭代中，后一个弱分类器训练的是前一个弱分类器的误差，且沿着最大下降梯度的方向。该模型是由Jerome

Friedman在1999年提出。作为一种迭代的决策树算法，该算法由多棵决策树组成，通常都是上百棵树，而且每棵树的规模都较小（即树的深度会比较浅通常为4到6）。模型预测的时候，对于输入的一个样本实例，首先会赋予一个初值，然后会遍历每一刻决策树，每棵树都会对预测值进行调整修正，最终的结果是将每一棵决策树的结果进行累加得到最后的预测的结果，如图1及公式（1）所示，其中β为权重系数[14-15]。

基于GBDT算法，可以更好地实现分类和回归任务，而且不容易出现过拟合现象。

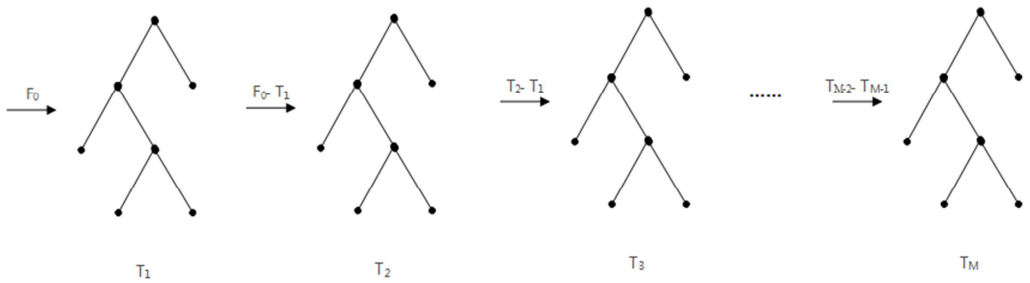


图1 GBDT算法思想示意图。

$$F(X)=F_0+\beta_1T_1(X)+\beta_2T_2(X)+...+\beta_MT_M(X) \quad (1)$$

3.4．数据处理软件

ECXEL 2010（Microsoft，USA）用于数据收集和整理；SIMCA-P11.5（Umetrics, Sweden）用于PCA、PLSDA及GBDT分类器分析。

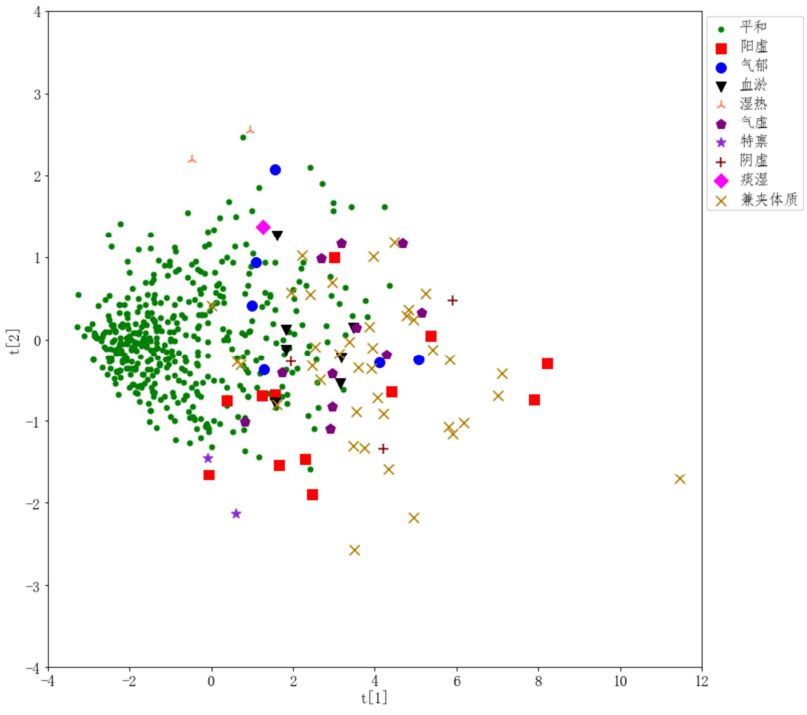
4．结果与分析

4.1．PCA 和 PLSDA 降维结果分析

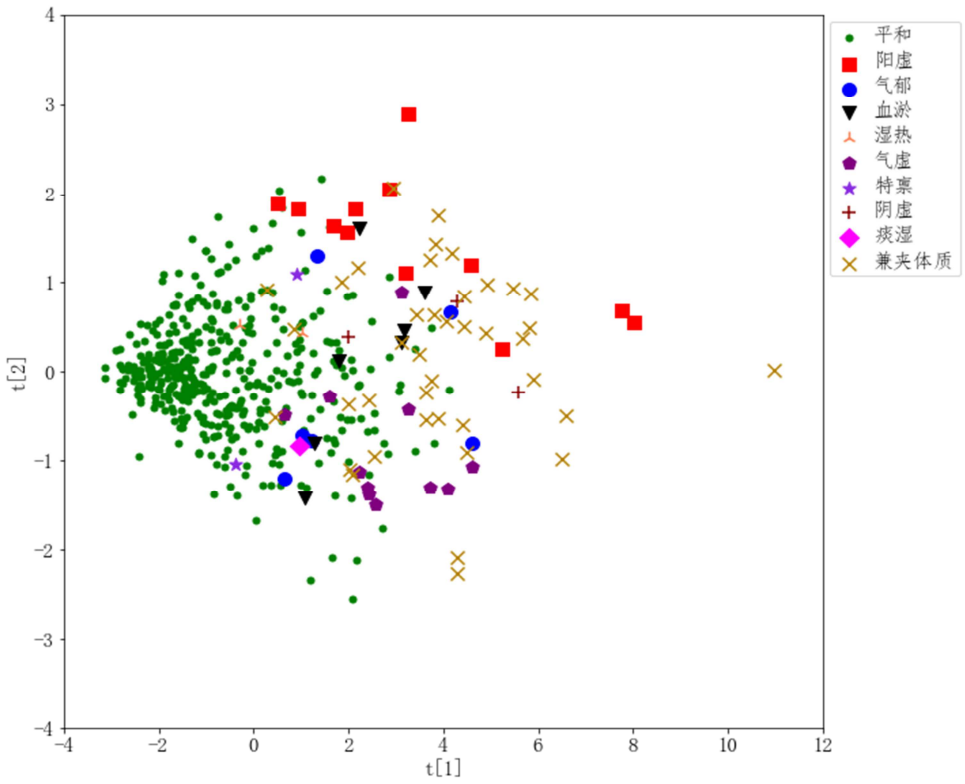
降维分析结果可以通过得分图和载荷图展示，结果如下：

（I）得分图

样本投影在两个主成分（t[1]和t[2]）构成的平面坐标系上，投影的得分值就是空间坐标，可以直观的反映样本的相似或差异性，如果两个样本之间相似性较高，那么两个坐标点在得分图上的位置相距较近，反之亦然。PCA和PLSDA的得分图分别如图2A、图2B。



A

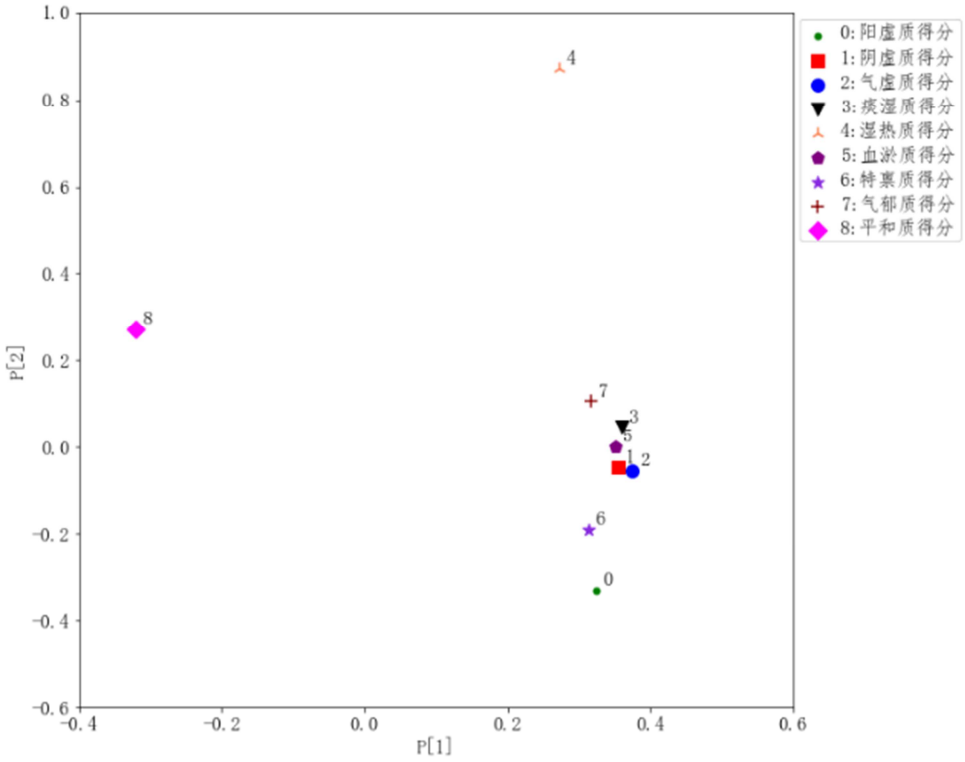


B

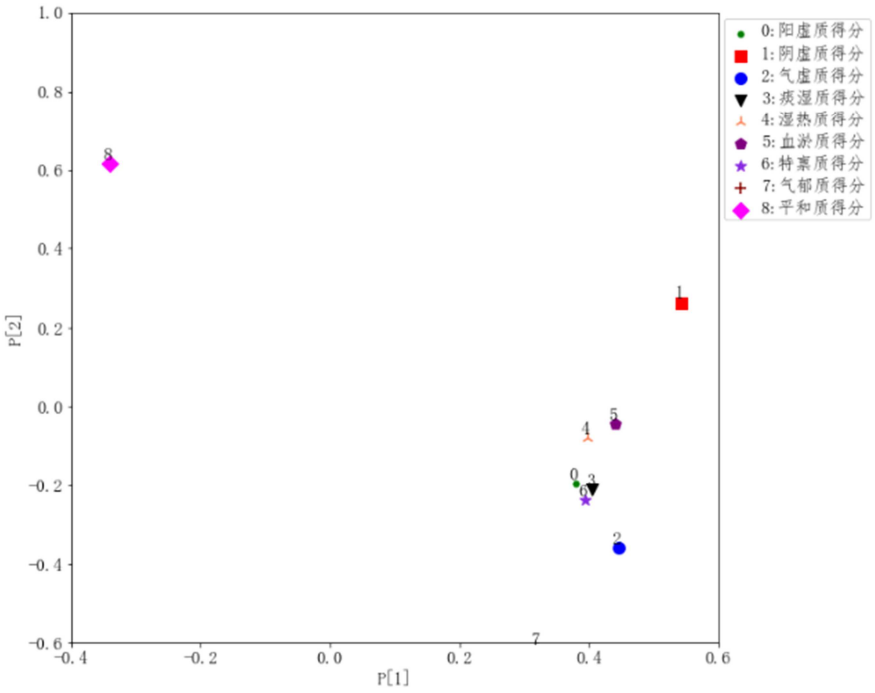
图2 PCA得分图及PLSDA得分图。

（II）荷载图

主成分载荷图可反应主成分与元变量之间的相互关联程度。如果原始特征中两个特征的相似性较高，则它们距离较近，反之亦然。PCA和PLSDA的荷载图分别如图3A、图3B。



A



B

图3 PCA荷载图及PLSDA荷载图。

4.2．PCA和PLSDA拟合参数对比

拟合参数R^2X常被用来评价模型的降维效果，这一指标反应的是模型降维后保留了原始数据信息量的多少。使用PCA和PLSDA对中医体质辨识数据进行降维，两种模型均保留2个主成分，得到PCA和PLSDA的拟合参数得分如表1：

表1 PCA和PLSDA的拟合参数得分。

R^2X	第一主成分（PC1）	第二主成分（PC2）
PCA	0.541	0.078
PLSDA	0.531	0.069

4.3．GBDT

如果使用了PCA、PLSDA降维后的分类效果跟原始特征的分类效果相差无几，则说明PCA、PLSDA降维之后，体质类型仍然能很好地通过这两个主成分表达出来。如果降维之后的分数与原始特征的分数相差较大，这说明降维的效果不好。使用PCA+GBDT 、PLSDA+GBDT和原始特征+GBDT在训练集（525个样本）训练后，在测试集（132个样本）中得到的分数如表2所示。

表2 PCA+GBDT 、PLSDA+GBDT和原始特征+GBDT所得分数。

	PCA+GBDT	PLSDA+GBDT	原始特征+GBDT
Accuracy score	0.878	0.841	0.871

5．讨论

PCA模型和PLSDA模型均能将复杂多元的体质评分结果降维得到一个二元的平面直观图。从两种模型的得分图及荷载图结果可知广东省佛山市顺德区龙江镇40-70岁年龄段的社区群众中平和质体质人群占大多数，兼夹体质人群次之。且平和质人群与其他体质类型人群的体质区别较大。另外，从PCA的荷载图还可得知，湿热质人群的体质与平和质及其他体质人群的体质区别也很大。为了更好地评价两种模型在中医体质分型中的效果及可靠性，我们引入了模型拟合参数及GBDT的概念来分别对其信息保留度及分类效果进行评估。

根据PCA模型和PLSDA模型的拟合参数得分，其中PCA模型的拟合参数R^2X，PC1=0.541，PC2=0.078，PLSDA模型的拟合参数R^2X，PC1=0.531，PC2=0.069。在PCA降维中，PC1、PC2的累积贡献率（2）$R^2X=(R^2X_{PC1}+R^2X_{PC2})$=0.619，即降到了两维后，PCA模型保留了原始数据61.9%的信息量。在PLSDA中，PC1、PC2的累计贡献率（3）$R^2X=(R^2X_{PC1}+R^2X_{PC2})$=0.6，即降到了两维后，PLSDA模型保留了原始数据60%的信息量。补充说明一下保留原始数据多少是可用的，再说PCA>PLSDA，故效果比它更好。

为更进一步阐述两种模型的分类效果，我们在中医体质辨识项目中，原始特征+GBDT的分类准确率为0.871（即在132个样本中预测准了132×0.878=115个）；使用PCA+GBDT处理后得到的分类准确率为0.878（即在132个样本中预测准了132×0.878=116）；使用PLSDA+GBDT的分类准确率为0.841（即在132个样本中预测准了132×0.841=111）。从数据结果可知PCA和PLSDA的分类准确率都很高，但PCA的分类更为准确。

6．结论

综合以上分析，我们可以得知PCA和PLSDA都可以对中医体质进行很好的归类，都能反应出平和质人群的体质

和其他体质人群相差很大。但是经过多种指标的检测发现，PCA的降维效果优于PLSDA。

基金项目

广州市高校创新创业教育项目（201709T15），广州市高校创新强校工程项目（Q17024006），深圳市光明区中医药科研项目（GM2019020001）

参考文献

[1] 杨靖，邢彤，李春禄．中医体质学说现代研究进展[J]．长春中医学院学报，2000，16(4)：601.

[2] 马晓慧．中医体质学说理论与方法研究进展[J]．上海中医药杂志，2000(5).

[3] 王琦．中医体质学[M]．中国医药科技出版社，1995．1.

[4] 何裕民，刘文龙．新编中医基础理论[M]．北京医科大学中国协和医科大学联合出版社，1996．113.

[5] 匡调元．两纲八要辨体质新论[J]．中医药学刊，2003．1(21).

[6] 赵蔷.主成分分析方法综述[J]．软件工程,2016,19(6):1-3.

[7] 刘彬球，陈孝权，吴晓刚，等. PCA和PLS-DA用于晒青毛茶级别分类研究[J]．茶叶科学,2015,35(02):179-184.

[8] 张威威，李瑞敏，谢中教．基于PCA-GBDT的城市道路旅行时间预测方法[J]．公路工程,2017,42(06):6-11.

[9] 杨天鸣，张璐，付海燕，等．不同产地甘草的近红外指纹图谱模式识别鉴别方法[J]．亚太传统医药,2015,11(14):11-14.

[10] 阮越，陈汉武，刘志昊，等．量子主成分分析算法[J]．计算机学报,2014,37(3):666-676.

[11] 陈佩．主成分分析法研究及其在特征提取中的应用[D]．陕西师范大学,2014.

[12] Alsberg B K, Kell D B, Goodacre R. Variable selection in discriminant partial least squares analysis. Analytical Chemistry, 1998, 70 (19): 4126-4133.

[13] Lutz U, Lutz W R, Lutz K W. Metabolic profiling of glucuronides in human urine by LC−MS/MS and partial least-squares discriminant analysis for classification and prediction of gender. Analytical Chemistry, 2006, 78 (13): 4564-4571.

[14] 王立平，邓芳明．基于小波包和GBDT的瓦斯传感器故障诊断[J]．测控技术，2016，35(12)：30—33.

[15] 郑凯文，杨超．基于迭代决策树(GBDT)短期负荷预测研究[J]．贵州电力技术,2017,20(02):82-84.

Biography

谢方镕(1993-)，女，硕士研究生，研究方向为消化内科中西医结合方向。

徐舒(1975-)，男，博士，主任医师，主要从事中西医结合消化疾病治疗与预防研究。