第30卷，第4期
2010年4月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 30, No. 4, pp984-987
April, 2010

# KPCA-聚类分析法和用便携式拉曼仪快速鉴别降糖药

翁欣欣[1]，张中湖[2]，尹利辉[3]，陆 峰[1*]

1. 第二军医大学，上海 200433
2. 山东省药品检验所，山东 济南 250012
3. 中国药品生物制品检定所，北京 100050

**摘 要** 对不同种类的降糖药片进行拉曼光谱的核主成分分析(KPCA)-聚类分析，实现快速、简便的鉴别。KPCA可以有效地避免主成分分析(PCA)只能处理线性问题和降维效果不明显的弊端。它通过一个非线性变换，首先将原变量空间映射到高维特征空间，然后在这个高维特征空间中进行线性主成分分析。采集得到的药片拉曼光谱的KPCA-聚类分析结果表明，采用KPCA提取特征变量的聚类结果比采用PCA提取特征变量后进行聚类分析的效果好，并且未经刮除表面包膜的降糖药片识别准确率为96.5%，经过刮除表面包膜处理的降糖药片的识别准确率为100%。便携式拉曼光谱仪结合该方法以其检测速度快、准确率高、使用简便、无样品前处理等显著优势，为药品的快速检验技术提供一种新的有效的鉴别手段。

**关键词** 便携式拉曼光谱仪；降糖药片；核主成分分析；聚类分析

**中图分类号**：R917 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2010)04-0984-04

## 引 言

药品现场抽验是药品监督管理工作中必要的技术支撑。每年国家用于药品抽验的经费达数亿元，因此药品快速检验已是当务之急[1-3]。药品现场快速检验，可以有效地避免药品的盲目抽验和无效抽验，降低药品检验成本。近年来，光谱分析方法发展迅速，其中拉曼光谱具有光谱信息丰富、非破坏性、耗时短、样品所需量小及样品无需制备等优点[4, 5]，在国内外的应用逐渐增多。便携式拉曼光谱仪更以其轻便、价廉等诸多优势，有望成为药品快速检验技术中一种新的鉴别手段[6-13]。核函数主成分分析(KPCA)是一种非线性方法，它在经典的主成分分析(PCA)的基础上，通过引入核函数将原空间中的非线性问题转化为映射空间中的线性问题[14]。本文采用便携式拉曼光谱仪，结合核主成分分析-聚类分析，考察其对部分降糖类药品(有些属于较难鉴别的化学结构与光谱信息相似的同系物)的识别能力。

## 1 实验部分

### 1.1 实验仪器

便携式拉曼光谱仪(由海洋光学中国分公司提供)，采用Laser 785-Lab激光器、Laphotonics 785 Probe探头与H6光栅。

### 1.2 实验样品及样品来源

样品分为刮除表面包膜(简称刮片)与未经刮除表面包膜(简称未刮片)。不同厂家、不同批次的盐酸二甲双胍片(未刮片1～12和刮片1～12)、格列齐特片(未刮片13～17和刮片13～18)、格列本脲片(未刮片18～23和刮片19～24)、格列吡嗪片(未刮片24～26和刮片25～27)和盐酸苯乙双胍片(未刮片27～29和刮片28～30)，由山东省药检所提供。

### 1.3 光谱采集

刮片与未刮片分别采用海洋光学便携式拉曼光谱仪扫描，扫描范围为2 028～0 $cm^{-1}$，狭缝50 $\mu$m，激发光源的起始波长为785 nm。

### 1.4 光谱信息处理

光谱数据预处理、PCA、KPCA、聚类分析等，均由MATLAB 7.0编程实现。

## 2 结果与讨论

### 2.1 光谱数据的预处理

采集的光谱样本受到高频随机噪声、基线漂移、光散射

**收稿日期**：2009-04-28，**修订日期**：2009-07-29

**基金项目**：国家科技支撑计划项目(2008BAI55B06)资助

**作者简介**：翁欣欣，女，1986年生，第二军医大学在读硕士研究生 e-mail：weng_xinxin@hotmail.com

*通讯联系人 e-mail：fenglufeng@hotmail.com

等影响，需要进行光谱预处理来消除噪声。选择光谱谱段 1 691.7～150.4 $cm^{-1}$，采用基线校正，5 点平滑，进行一阶导数预处理，然后进行向量标准化，得到的光谱先进行特征变量提取。

### 2.2　主成分分析得到新的特征变量

PCA 把多个指标化为少数几个综合指标，其主要目的是降维[15, 16]。经主成分分析光谱数据后，得到前 10 个主成分累积贡献率如表 1、表 2 所示，前 6 个主成分的累积贡献率已经达到了 98.663%和 98.319%，能够比较全面的反映所有信息，所以每个样品的光谱数据可以用 6 个主成分代替。

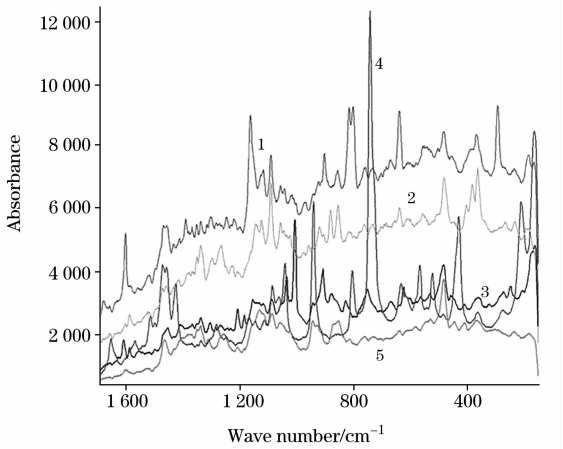


**Fig. 1　Raman spectrum of gliclazide**

1: Gliclazide tablet; 2: Glipizide tablet; 3: Phenformin Hydrochloride tablet; 4: Metformin Hydrochloride tablet; 5: Glibenclamide tablet

**Table 1　10PCs and reliabilities (with coating)**

| 主成分 | 累积贡献率/% | 主成分 | 累积贡献率/% |
|---|---|---|---|
| PC1 | 52.852 | PC6 | 98.663 |
| PC2 | 74.008 | PC7 | 99.084 |
| PC3 | 86.486 | PC8 | 99.323 |
| PC4 | 95.127 | PC9 | 99.446 |
| PC5 | 97.952 | PC10 | 99.541 |

**Table 2　10PCs and reliabilities (without coating)**

| 主成分 | 累积贡献率/% | 主成分 | 累积贡献率/% |
|---|---|---|---|
| PC1 | 49.128 | PC6 | 98.319 |
| PC2 | 70.933 | PC7 | 98.791 |
| PC3 | 84.026 | PC8 | 99.112 |
| PC4 | 92.772 | PC9 | 99.373 |
| PC5 | 96.751 | PC10 | 99.578 |

### 2.3　核主成分分析得到新的特征变量

KPCA 通过引入核函数把数据非线性映射到高维核空间，在核空间利用传统的 PCA 技术进行特征提取，描述了高维特征间的相关性。KPCA 的基本思想就是将一个非线性映射 $\Phi$ 作用于输入空间 $R^N$ 后，使得在特征空间中的问题变得线性可分。

经核主成分分析光谱数据后，得到前 10 个主成分累积贡献率如表 3、表 4 所示，前 4 个主成分的累积贡献率已经达到了 99.119%和 97.911%，所以每个样品的光谱数据可以用 4 个主成分代替。

**Table 3　10PCs and reliabilities (with coating)**

| 主成分 | 累积贡献率/% | 主成分 | 累积贡献率/% |
|---|---|---|---|
| PC1 | 73.850 | PC6 | 99.887 |
| PC2 | 89.165 | PC7 | 99.925 |
| PC3 | 96.347 | PC8 | 99.953 |
| PC4 | 99.119 | PC9 | 99.976 |
| PC5 | 99.788 | PC10 | 99.991 |

**Table 4　10PCs and reliabilities (without coating)**

| 主成分 | 累积贡献率/% | 主成分 | 累积贡献率/% |
|---|---|---|---|
| PC1 | 68.041 | PC6 | 99.285 |
| PC2 | 86.544 | PC7 | 99.585 |
| PC3 | 95.159 | PC8 | 99.741 |
| PC4 | 97.911 | PC9 | 99.816 |
| PC5 | 98.836 | PC10 | 99.875 |

### 2.4　聚类分析

系统聚类法是目前最常用的一种聚类方法。本文以标准欧氏距离作为衡量各种降糖药种类差异大小，采用最小方差算法对降糖药进行系统聚类。

图2和图3为没有经过特征变量提取的聚类分析结果

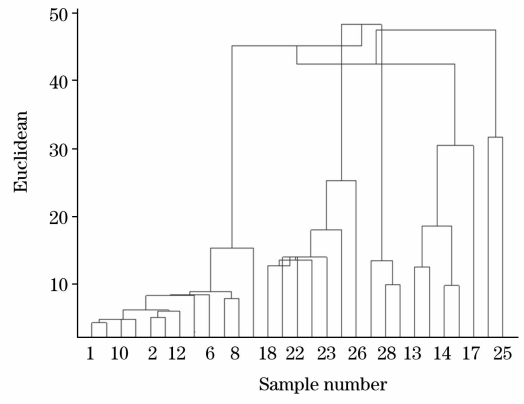


**Fig. 2　Dendrogram of tablets (with coating)**

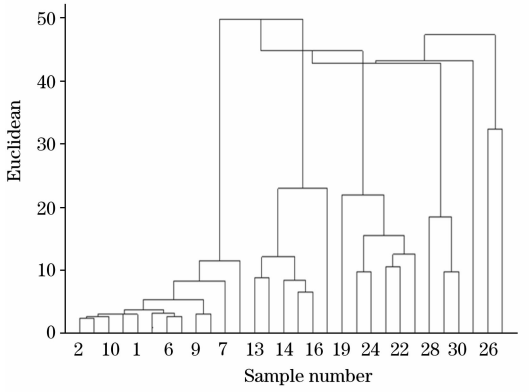


**Fig. 3　Dendrogram of tablets (without coating)**

图，图 4 和图 5 为经过 PCA 提取特征变量后的聚类分析结果图，图 6 和图 7 为经过 KPCA 提取特征变量后的聚类分析结果图。由图 2 和图 3 所示，将拉曼数据直接进行聚类分析，无法对药片进行准确分类。由图 4 和图 5 所示，未刮片中只有盐酸二甲双胍片被正确区分，而刮片中盐酸二甲双胍片和格列本脲片被正确区分，但是其他 3 种药片依然没有正确的分类。由图 6 和图 7 所示，采用 KPCA-聚类分析方法时，未刮片除了 26 号药片被分类错误，其他均正确区分，刮片的聚类分析结果显示所有的药片均准确区分。

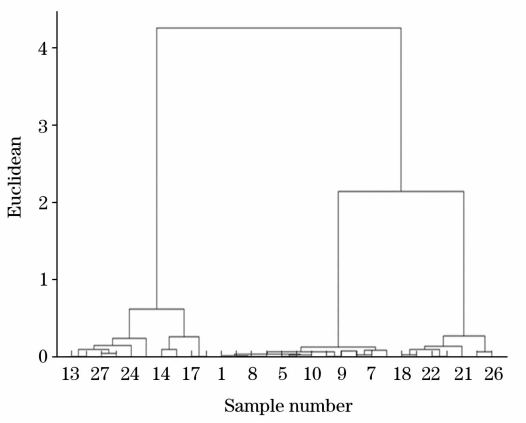


**Fig. 4 Dendrogram of tablets (with coating)**

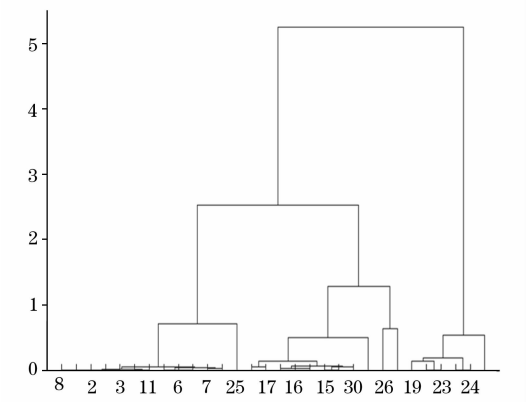


**Fig. 5 Dendrogram of tablets (without coating)**

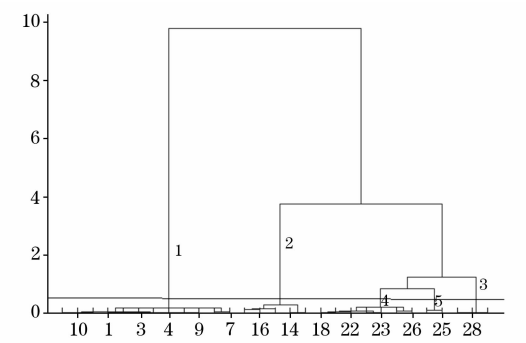


**Fig. 6 Dendrogram of tablets (with coating)**

### 2.5 讨 论

(1)在模式分类问题中，过高的数据维数不仅会增加数据采集难度，还会增加存储空间及数据噪声，从而影响分类器的性能。特征选择的任务就是从初步选出的特征集中进一步筛选能实现分类性能最大化的最小特征子集，以降低数据采集难度和数据噪声。在提取特征变量时，KPCA 比 PCA 的降维效果好。另外，PCA 是用来处理变量间线性关系的，而 KPCA 不仅能处理变量间的线性关系，而且能有效地处理非线性关系。因此核主成分分析对分类器的性能改善效果更好，采用 KPCA 提取特征变量的聚类结果比采用 PCA 提取特征变量后进行聚类分析的总体效果好。

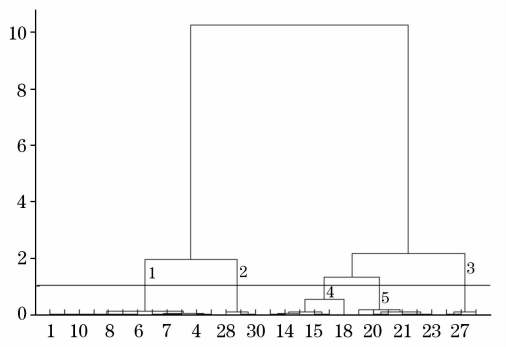


**Fig. 7 Dendrogram of tablets (without coating)**

(2)药片拉曼光谱的 KPCA-聚类分析结果表明，未刮片的识别准确率为 96.5%，刮片的识别准确率则达到了 100%。未刮片由于包膜信息与片芯成分信息重叠的影响，使得药片中化合物的骨架振动信号相对减弱，从而影响了分类的准确性。尽管如此，由于拉曼光谱反映的是化合物的骨架振动，特征性强，片芯成分与包膜成分的光谱信息特征仍有较大的差异，即使是格列齐特、格列本脲、格列吡嗪等属于同系物，其骨架振动上的细微差异导致了一定的光谱特征差异(见图 1)，结合 KPCA-聚类分析，仍能得到较好的识别效果。

(3) 有文献报道[7]采用建立药品相应的对照品谱库，通过识别出主成分的方法判定分析药品的种类。但是从药品的拉曼光谱中识别出主成分，受到其含量大小、拉曼特征强弱以及辅料干扰等因素的制约，这些因素会导致主成分识别比较困难，从而影响药品种类的正确区分。本文采用 KPCA-聚类分析法，直接对降糖药进行分类，无需建库，方便准确。

(4) 由于药片大部分是非均匀相，会导致拉曼光谱数据采集时发生较大变化。因此如何从样品处获取稳定的光谱以提高分析判定的准确可靠性，有待进一步研究。综合上述结果，便携式拉曼光谱仪结合 KPCA 提取特征变量的聚类分析可以用于降糖药片的区分鉴别。

## 3 结 论

便携式拉曼光谱仪体积小，轻巧便捷，不需要其他辅助测量装备即可完成检测。KPCA-聚类分析法可以对不同的降糖药片进行快速鉴别，并且具有很强的识别能力。因此，拉曼光谱与 KPCA-聚类分析法相结合能无损快速鉴别降糖药片的种类。

## 参　考　文　献


[1] ZHANG Chun-bo, FAN Yu-feng(张春波, 范玉峰). China Medical Herald(中国医药导报), 2008, 5(28): 109.
[2] LI Shu, CAO Yan, LE Jian, et al(李　树, 曹　岩, 乐　健, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2009, 29(2): 327.
[3] Fernandez Facundo M, Green Michael D, Newton Paul N. Industrial and Engineering Chemistry Research, 2008, 47: 585.
[4] Kudelski Andrzej. Talanta, 2008, 76: 1.
[5] WANG Yu, LI Zhong-hong, ZHANG Zheng-xing, et al(王　玉, 李忠红, 张正行, 等). Acta Pharmaceutica Sinica(药学学报), 2004, 39(9): 764.
[6] Pesce William J, Wiley Peter Booth. Pharmaceutical Applications of Raman Specuroscopy. Slobodan Šašic, 2007.
[7] ZHANG Xin, LIU Zhao-xia, NI Kun-yi, et al(张　新, 刘朝霞, 倪坤仪, 等). Chinese Pharmaceutical Affairs(中国药事), 2008, 22(7): 555.
[8] de Peinder P, Vredenbregt M J, Visserc T, et al. Journal of Pharmaceutical and Biomedical Analysis, 2008, 47: 688.
[9] Eliasson Charlotte, Macleod Neil A, Jayes Linda C, et al. Journal of Pharmaceutical and Biomedical Analysis, 2008, 47: 221.
[10] de Veij Marleen, Deneckere Annelien, Vandenabeele Peter, et al. Journal of Pharmaceutical and Biomedical Analysis, 2008, 46: 303.
[11] Ricci Camilla, Nyadong Leonard, Yang Felicia, et al. Analytica Chimica Acta, 2008, 623: 178.
[12] de Veij Marleen, Vandenabeele Peter, Hall Krystyn Alter, et al. Journal of Raman Spectroscopy, 2007, 38: 181.
[13] ZHANG Qi-ming, ZHANG Xin, LIU Zhao-xia(张启明, 张　新, 刘朝霞). Chinese Pharmaceutical Journal(中国药学杂志), 2008, 43(24): 1903.
[14] Xu Yong, Zhang David, Song Fengxi, et al. Neurocomputing, 2007, 70: 1056.
[15] LIANG Yi-zeng(梁逸曾). White, Grey and Black Multicomponent Systems and Their Chemometric Algorithms(白灰黑复杂多组份分析体系及其化学计量学算法). Changsha: Hunan Publishing House of Science and Technology(长沙: 湖南科学技术出版社), 1996.
[16] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Chinese Journal of Analytical Chemistry(分析化学), 2000, 28(4): 421.


# Rapid Determination of Hypoglycemic Tablets by Handheld Raman Spectrometer and KPCA-Clustering Analysis


WENG Xin-xin[1], ZHANG Zhong-hu[2], YIN Li-hui[3], LU Feng[1]*

1. School of Pharmacy, Second Military Medical University, Shanghai　200433, China
2. Shandong Provincial Institute for Drug Control, Ji'nan　250012, China
3. National Institute for the Control of Pharmaceutical and Biological Products, Beijing　100050, China



**Abstract**　In the present paper, five different kinds of hypoglycemic tablets were identified using kernel principal component analysis (KPCA)-clustering analysis of their Raman spectra. KPCA was used to compress thousands of spectral data into several variables and to describe the body of the spectra before clustering analysis was chosen as further research method. The results showed that hypoglycemic tablets could be quickly classified using KPCA-clustering analysis. A disadvantage of Raman spectroscopy for this type of analysis is that it is primarily a surface technique. As a consequence, the spectra of the tablet core and its coating might differ. However, the KPCA-clustering analysis turned out to be a sufficiently reliable discrimination, i.e., 96% of the hypoglycemic tablets with coating and 100% of the hypoglycemic tablets without coating were predicted correctly. Overall, the Raman spectroscopic method in the present paper plays a good role in the identification and offers a new approach to the rapid discrimination of different kinds of hypoglycemic tablets.



**Keywords**　Handheld Raman spectrometer; Hypoglycemic tablets; Kernel principal component; Clustering analysis



* Corresponding author



(Received Apr. 28, 2009; accepted Jul. 29, 2009)