

Identification of Specific Liquor Based on Near-Infrared Spectroscopy Technology

Tong Wu*, Chao Tan

Key Lab of Process Analysis and Control of Sichuan Universities, Yibin University, Yibin Sichuan
Email: *wutmail@163.com

Received: Jan. 16th, 2016; accepted: Feb. 1st, 2016; published: Feb. 5th, 2016

Copyright © 2016 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

A rapid nondestructive method for kind identification of liquor by near infrared spectroscopy (NIR) was investigated and developed. A total of 100 samples belonging to 8 brands were collected in local markets. Based on the wavenumber region of 5340-7400 cm^{-1} , four kinds of algorithm, *i.e.*, K-nearest neighbor's method of all spectral variables (KNN1) and of the first four principal components (KNN2), perception of the first four principal components and Fisher linear discrimination of all spectral variables, were used for constructing the classification models. By comparison, it shows that the Fisher linear discrimination model can achieve the best performance, with a misclassified ratio of 0% in either the training set or the test set. The results also indicate that the combination of NIR and pattern recognition models is feasible for a rapid identification of liquor such as Wuliangye.

Keywords

Near-Infrared Spectroscopy, Liquor, Identification, Fisher Linear Discriminant

基于近红外光谱技术快速鉴别白酒真伪

吴同*, 谭超

宜宾学院过程分析与控制四川省高校重点实验室, 四川 宜宾
Email: *wutmail@163.com

收稿日期: 2016年1月16日; 录用日期: 2016年2月1日; 发布日期: 2016年2月5日

*通讯作者。

文章引用: 吴同, 谭超. 基于近红外光谱技术快速鉴别白酒真伪[J]. 分析化学进展, 2016, 6(1): 1-6.
<http://dx.doi.org/10.12677/aac.2016.61001>

摘要

本文探索了一种利用近红外光谱技术进行白酒真伪快速鉴别的新方法。共收集到分属8个品牌的100个酒样, 基于 $5340\sim 7400\text{ cm}^{-1}$ 的光谱区域, 使用四种化学计量学方法建立分类模型, 即基于全部变量的K最近邻模型、基于前四个主成分得分的K最近邻模型, 基于前四个主成分得分的感知机、基于全部变量的Fisher判别模型。比较显示, Fisher判别模型达到了最优的性能, 在训练集和测试集上的误判率均为0%。可见近红外光谱技术结合模式识别方法可用于快速鉴别五粮液等高端酒真伪。

关键词

近红外光谱, 白酒, 鉴别, Fisher判别法

1. 引言

中国白酒是有上千年悠久历史的传统产业。从组成上看, 白酒的主要成分是乙醇和水, 约占白酒总量的98%~99%, 而微量香味成分含量只占白酒总量1%~2%, 而恰好是这些难于检测的微量成分一起决定了白酒的品质[1]。这些成分及其含量对某一个品牌是具有专一性, 是保持该品牌质量稳定以及区别于其他品牌的关键, 并将直接影响消费者对品牌的选择。白酒, 尤其是茅台、五粮液等名酒, 多年来一直受假名酒的侵害。假冒伪劣白酒不仅严重影响了白酒的信誉、扰乱了市场的正常秩序, 也严重损害了消费者的切身利益和身体健康。如何区分和鉴别特定名优白酒的真伪, 以保护生产企业的利益和消费者的健康, 十分必要。

长期以来, 白酒的分析和评价一直使用感官检验, 即利用人的感觉器官——眼、鼻、口来综合判断酒的色、香、味和风格, 其具有简便、快捷的优点, 且直接与人的喜好相联系。然而, 感官评定不够科学、规范, 而且不客观, 常会受到主客观因素和条件的影响, 而显失偏颇。20世纪60年代以后, 现代仪器分析技术特别是色谱技术, 在被引入白酒生产和检测领域之后, 给白酒的分析和评价带来了深远的影响。研究发现: 各香型的名优白酒中所含的香味成分种类基本相同的, 但其量高低及它们之间的量比关系则不同。试图对白酒中的数百种化学成分进行逐一分析定量既无必要也不可能。色谱法属于分离检测方法, 存在分析耗时长、操作条件不易控制等问题。

近红外光谱是当前迅速发展的一项测试技术, 具有快速、无损耗、多成分同时分析、分析过程无污染高等优点, 在农业、食品、化工、制药、纺织、烟草等领域得到了成功应用[2]-[8]。仪器的近红外信号主要对应于分子含氢基团的合频吸收与倍频吸收, 可以反映出有机物的大量信息。由于干扰和背景影响, 近红外光谱具有谱峰重叠、变动、信号弱的特点。如何借助化学计量学方法从复杂、重叠、变动的光谱中提取特征信息, 建立定性定量模型是制约近红外光谱技术发展和应用的难题。

2. 实验部分

2.1. 样品准备

实验样品均为市售, 在不同的局部市场总共收集不同批次8个品牌的100个瓶装浓香白酒样品, 其中五粮液酒样30个, 五粮春28个、古蔺6个、金六福8个、老梦酒8个、绵竹大曲8个、南福大曲6个、叙府大曲6个。

2.2. 光谱采集

本实验采用美国ThermoFisher公司生产的Antaris II傅立叶变换近红外光谱仪, 采用光纤探头透反射

模式进行测量, 光程为 2 mm, 采集范围为 4000~10,000 cm^{-1} , 扫描次数为 16, 光谱分辨率为 4 cm^{-1} , 每个光谱包含 1537 个点。采用配套的 RESULT 软件辅助编写工作流程和控制采集。

2.3. 化学计量学方法

采用 K 最近邻法、感知机、Fisher 线性判别分类器作为模型建立方法, 以下是方法的简要介绍。

K 最近邻法[9] (K-nearest neighbors, 简称 KNN 法)是一种直接以模式识别的基本假设“同类样本在模式空间相互靠近”为依据的非参数分类方法。每一个样本可以看作特征或变量空间里该空间中的一个点。对于一未知样本, 可通过分析与它距离最近的一些已知样本点的类别及其数量加以分类, 此方不受体系线性可分与否的限制, 将代表性训练样本数据存储在计算机上, 对待判别的未知样本, 计算其与各训练集样本间的距离, 找出距离未知样本点最近的 k 个训练样本点, 然后进行判别。目前尚未有指导性的 k 取值原则, 一般通过试验确定。

感知机是一种常见的神经网络分类方法[10], Frank Rosenblatt 等人 20 世纪 50 年代末提出, 感知机学习旨在求出将训练数据集进行线性划分的分类超平面, 感知机学习算法是误分类驱动的, 具体采用随机梯度下降法先导入基于误分类的损失函数, 然后利用梯度下降法对损失函数进行极小化, 从而求出感知机模型。感知机模型是神经网络和支持向量机的基础。感知机网络结构简单, 学习算法易于实现。

Fisher 线性判别分析[11] [12]是较常使用一类判别分析方法。Fisher 判别是依据方差分析原理建立起来的, 其基本思路就是投影, 针对 p 维空间中的某点寻找一个能使它降为一维数值的线性函数。然后应用这个线性函数把 p 维空间中的训练集样本点以及未知归属的测试样本都变换为一维数据, 再根据其间的亲疏程度把未知归属的样本点判定其归属。这个线性函数应能在把 p 维空间中的点转化为一维数值后, 既能最大限度地缩小同类样本点之间的差异, 又能最大限度地扩大不同类样本点之间的差异, 这样才可能获得较高的判别效率。

3. 结果与讨论

总共 100 个酒样, 按照前述的光谱采集方法, 获得 100 个光谱, 每个光谱分配一个类别号。五粮液样单独作为一类, 分配类别号 1, 其余 70 个不同价位的中低端酒样视为其余类(非五粮液), 分配类别号 2。

对于一个基于光谱的分类问题, 一般要先将其划分为一个训练集和一个测试集。为了保证模型验证的可靠性, 使两个子集有尽可能相似的信息分布。随机均分各类酒样, 使得训练集和测试集均含 50 个样本。图 1 给出了所有的原始光谱。由于光谱在某些区域存在的吸收饱和现象, 试验挑选了 5340~7400 cm^{-1} 区域用于分类识别。图 1 也给出了各类酒样在所选区间的平均谱, 可见, 各类酒样的近红外光谱十分相似, 肉眼难于分辨。

近红外光谱存在严重的共线性和重叠, 且数据点多, 为揭露光谱数据的内部结构和可能的样本分布规律, 对数据进行主成分分析(PCA), 将原变量进行转换, 用较少的新变量(主成分)浓缩原变量信息。图 2 给出了累计解释的方差随主成分数的变化。可见, 前三个主成分已能够解释 99% 以上的数据方差。进一步, 图 3 显示了所有样本主成分得分散点图。可见, 五粮液样本点分布相对集中, 而其余几类酒的样本点分布在其周围, 也能大致看出, 至少在二维空间(PC1-PC2)是线性不可分。

共采用四种方法来建立分类模型, 即基于全部变量的 K 最近邻模型(KNN1)、基于前四个主成分得分的 K 最近邻模型(KNN2), 基于前四个主成分得分的感知机(Perceptron)、基于全部变量的 Fisher 线性判别模型(Fisher), 其中的主成分数通过试验先确定。图 4 显示了四类模型的分在测试集上的性能比较。可见, 采用 K 最近邻模型, 无论是基于全部变量还是四个主成分得分, 分类效果相同, 错分的样本号也一

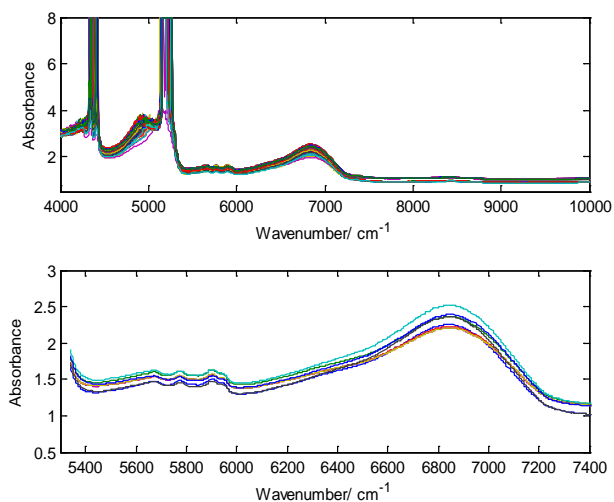


Figure 1. Original spectra and the average spectra in selected range

图 1. 原始光谱和各类在所选区间的平均谱

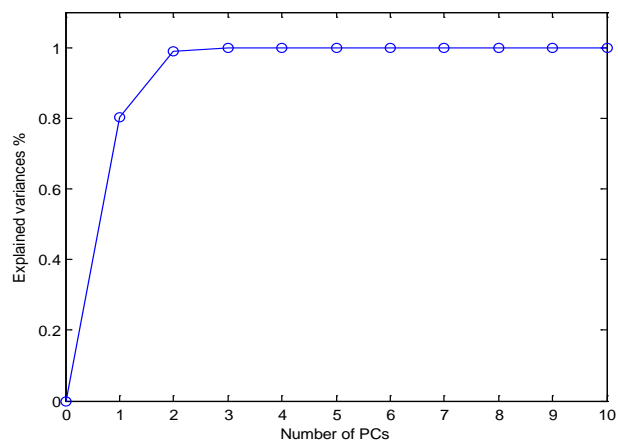


Figure 2. Change of explained variance with the number of principal components

图 2. 解释的方差随主成分数的变化

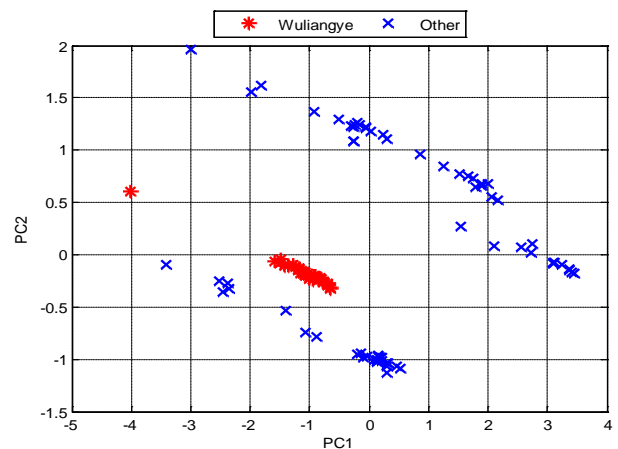


Figure 3. Scatter plot of the scores of principal components

图 3. 主成分得分散点图

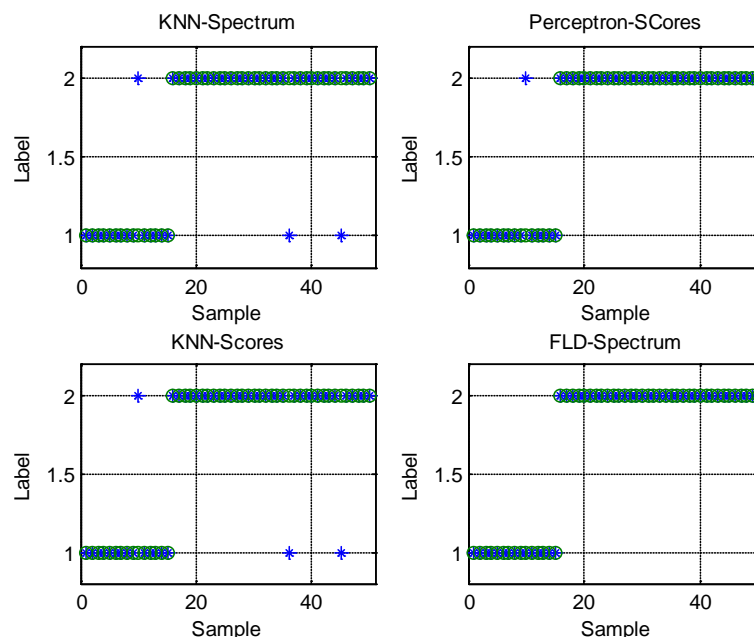


Figure 4. Comparison of the classification performance of four models
图 4. 四类模型的分类型性能比较

Table 1. Comparison of the failure rate of four models
表 1. 四类模型误判率指标比较

	KNN1	KNN1	Perceptron	FLD
Training set	0.02	0.02	0	0
Test set	0.06	0.06	0.02	0

致。而对基于前四个主成分得分的感知机模型，仅 1 个五粮液样被错分，相比 K 最近邻模型，感知机模型性能有一定提升，准确率更高。而对于基于全部变量的 Fisher 线性判别模型，则全部测试样品均分类正确。

表 1 给出了四类模型误判率指标，可见，对于四类模型，在测试集上的误判率均大于在训练集上的误判率，对于感知机和 Fisher 模型，所有的训练集样本都被正确识别；Fisher 模型的训练集和测试集上的误判率均为 0，说明所有的样本均能够正确识别，是最佳模型，模型对训练集的拟合能力好对独立测试集的预测性能也很好，无过拟和出现。

4. 结论

本文联合近红外光谱技术与分类模型对五粮液酒的真伪鉴别进行了初步研究。结果表明，5340~7400 cm^{-1} 区域光谱信号组合主成分感知机模型和 Fisher 判别模型能达到较好的识别率。两种方法所建模型在训练集上的误判率均为 0%，在测试集上的误判率率分别为 2%和 0%。结果表明近红外光谱技术结合模式识别方法快速鉴别五粮液等高端酒的真伪是可行的，为白酒的质量控制及市场监管检测提供了一种新方法。尽管只针五粮液等几类酒进行了实验，但实验结果具有普遍意义，可用于其它类似项目检测。将来可通过收集更多的、覆盖面更广的训练集样品来建立更可靠的模型。

基金项目

感谢国家自然科学基金 (21375118)和发酵资源与应用四川省高校重点实验室(2011KFJ001)的资助。

参考文献 (References)

- [1] Saurina, J. (2010) Characterization of Wines Using Compositional Profiles and Chemometrics. *Trends in Analytical Chemistry*, **29**, 234-245. <http://dx.doi.org/10.1016/j.trac.2009.11.008>
- [2] Song, H.Y., Qin, G. and Liu, H.Q. (2012) Qualitative Detection of Bottled Vinegar Based on NIR Spectroscopy Technique. *Spectroscopy and Spectral Analysis*, **32**, 1547-1549.
- [3] Cozzolino, D., Kwiatkowski, M.J., Parker, M., *et al.* (2004) Prediction of Phenolic Compounds in Red Wine Fermentations by Visible and Near Infrared Spectroscopy. *Analytica Chimica Acta*, **513**, 73-80. <http://dx.doi.org/10.1016/j.aca.2003.08.066>
- [4] Tan, C., Wu, T., Xu, Z.H., *et al.* (2012) A Simple Ensemble Strategy of Uninformative Variable Elimination and Partial Least-Squares for Near-Infrared Spectroscopic Calibration of Pharmaceutical Products. *Vibrational Spectroscopy*, **58**, 44-49. <http://dx.doi.org/10.1016/j.vibspec.2011.09.011>
- [5] Laxalde, J., Ruckebusch, C., Devos, O., *et al.* (2011) Characterisation of Heavy Oils Using Near-Infrared Spectroscopy: Optimisation of Pre-Processing Methods and Variable Selection. *Analytica Chimica Acta*, **705**, 227-233. <http://dx.doi.org/10.1016/j.aca.2011.05.048>
- [6] Jing, M., Cai, W.S. and Shao, X.G. (2010) Multiblock Partial Least Squares Regression Based on Wavelet Transform for Quantitative Analysis of Near Infrared Spectra. *Chemometrics and Intelligent Systems*, **100**, 22-27. <http://dx.doi.org/10.1016/j.chemolab.2009.09.006>
- [7] Liu, F., Ye, X.J., He, Y., *et al.* (2009) Application of Visible/Near Infrared Spectroscopy and Chemometric Calibrations for Variety Discrimination of Instant Milk Teas. *Journal of Food Engineering*, **93**, 127-133. <http://dx.doi.org/10.1016/j.jfoodeng.2009.01.004>
- [8] Tan, C., Chen, H., Wu, T., *et al.* (2013) Determination of Total Sugar in Tobacco by Near-infrared Spectroscopy and Wavelet Transformation-Based Calibration. *Analytical Letters*, **46**, 171-183. <http://dx.doi.org/10.1080/00032719.2012.704538>
- [9] Ni, L.J., Zhang, L.G., Xie, J., *et al.* (2009) Pattern Recognition of Chinese Flue-Cured Tobaccos by an Improved and Simplified K-Nearest Neighbors Classification Algorithm on Near Infrared Spectra. *Analytica Chimica Acta*, **633**, 43-50. <http://dx.doi.org/10.1016/j.aca.2008.11.044>
- [10] Nahum, M., Daikhin, L., Lubin, Y., *et al.* (2010) From Comparison to Classification: A Cortical Tool for Boosting Perception. *The Journal of Neuroscience*, **30**, 1128-1136. <http://dx.doi.org/10.1523/JNEUROSCI.1781-09.2010>
- [11] Chiang, L.H., Kotanchek, M.E. and Kordon, A.K. (2004) Fault Diagnosis Based on Fisher Discriminant Analysis and Support Vector Machines. *Computers and Chemical Engineering*, **28**, 1389-1401. <http://dx.doi.org/10.1016/j.compchemeng.2003.10.002>
- [12] Liang, Z.Z. and Shi, P.F. (2004) An Efficient and Effective Method to Solve Kernel Fisher Discriminant Analysis. *Neurocomputing*, **61**, 485-493. <http://dx.doi.org/10.1016/j.neucom.2004.06.005>