

基于激光光致发光光谱的发票真伪鉴别方法研究

杨琴¹, 杨勇², 田永红¹

1. 长江大学物理科学与技术学院, 湖北 荆州 434023
2. 中国地质大学数学与物理学院, 湖北 武汉 430074

摘要 用激光光致发光光谱技术对发票真伪的快速鉴别进行了研究。首先采集 80 个真假发票样本的激光光致发光光谱数据, 然后利用遗传算法对谱线重叠严重的 566~669 nm 波段进行高斯拟合, 提取特征参数, 再结合 BP 神经网络技术建立鉴别模型。结果表明, 从 104 个光谱数据提取 13 个特征参数, 真假发票的拟合确定系数 R^2 分别为 0.997 89 和 0.996 83, 相对标准差 RSD 分别为 0.017 052 和 0.022 362, 有效实现对原始光谱特征信息的提取和简化; 将 13 个特征参数和 2 个评估参数 R^2 , RSD 作为 BP 神经网络的输入变量, 模型质量最高, 对 60 个建模发票样本和 20 个未知发票样本的识别准确率均达到 100%。说明文章提出的光谱分析方法具有很好的分类和鉴别作用, 为发票真伪的快速鉴别提供了一种新方法。

关键词 激光光致发光; 发票; 遗传算法; 高斯拟合; BP 神经网络; 真伪鉴别

中图分类号: O433; F810.42 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2011)12-3328-04

引言

发票作为记录经营活动的一种原始凭证, 不仅是财务收支的法定凭证, 也是税务稽查的重要依据, 常被称为第二钞票。目前, 我国普通发票的管理机制存在严重的缺陷和漏洞, 防伪科技含量不高, 极易被仿制。据国家税务总局的不完全统计, 全国税务机关每年向纳税人发放的普通发票在 150 亿到 200 亿之间, 许多违法分子制售和使用假发票渐成泛滥之势, 给国家财政收入造成巨大损失。因此, 不断提高发票真伪鉴别的技术手段, 已成为我国税务部门面临的需迫切解决的问题, 意义深远。

税务部门在发票的防伪上采取了相当多的手段, 如使用防伪油墨、水印发票纸、电话防伪、密码防伪等^[1]。由于发票的监管机制尚未健全, 又考虑防伪技术成本, 充分利用现行印刷工艺, 一般考虑用荧光油墨套印发票监制章及发票号码, 即荧光防伪技术^[2]。然而, 在室温下防伪油墨的荧光光谱具有宽带结构, 谱带的重叠现象非常严重; 在违法犯罪现场查获的假发票大都制作精良, 发票监制章及发票号码也均使用防伪油墨, 只因成分不同导致在外来光照射下产生的荧光颜色会稍有不同, 获得的荧光光谱峰的峰形和峰位也只有在细微的差别, 仿真度极高以致真伪难辨。

光致发光 (photoluminescence, PL) 是指物质依赖外界光源进行照射, 从而获得能量产生激发导致发光的现象。本工作采用快速、无损、灵敏度及选择性都很高的激光光致发光光谱技术^[3,4], 对由湖北省武汉市监制的普通发票进行了真伪鉴别研究。利用绿光半导体激光照射真假发票的监制章发射荧光, 通过全局搜索算法-遗传算法^[5,6] (genetic algorithm, GA) 对真假发票的光谱曲线进行高斯拟合, 提取特征参数, 作为 BP (back propagation) 神经网络^[7] 的输入变量, 发票真伪作为神经网络的输出变量, 建立神经网络模型, 对普通发票的真伪进行快速鉴别。

1 材料与方法

1.1 仪器设备

实验使用北京卓立汉光公司的 ZLX-PL-I 型光致发光光谱测量系统, 如图 1 所示。光源采用全固体化单频绿光激光器, 输出波长 532 nm, 输出功率 > 200 mW; Omni- λ 300 型光谱仪的扫描范围 0~1 200 nm, 分辨率 0.1 nm; R2949 型光电倍增管 (PMT) 的光谱响应范围 185~900 nm, 峰值响应波长 (400 \pm 30) nm; 高压稳压电源 (HV) 0~1 500 V 连续可调; 美国 SRS 公司的 SR830 型锁相放大器 (PLA); 数据采集器 (DCS) 等。

收稿日期: 2011-03-10, 修订日期: 2011-07-16

基金项目: 国家自然科学基金项目 (49503046, 11047176) 和湖北省高等学校科研基金项目 (2002B04002) 资助

作者简介: 杨琴, 女, 1982 年生, 长江大学物理科学与技术学院讲师 e-mail: yangqin@yangtzeu.edu.cn

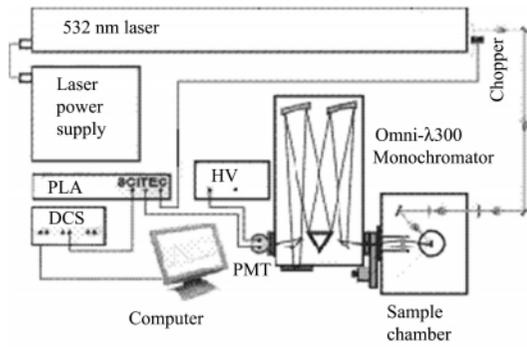


Fig 1 General structural diagram of laser-induced photoluminescence detection system

1.2 样品制备及光谱的获取

从由湖北省武汉市监制的真发票和市场上抽检出的假发票中各挑选 40 张, 实验前分别置于 5 kg 重物下 10 h, 以消除发票表面不平对其光谱特征的影响。从全部 80 个样本中, 真假发票各随机选择 30 个共 60 个样本作为建模集, 剩余的各 10 个共 20 个样本作为预测集。系统经校准后进行测试: 激光束直径约 0.5 mm, 激光器电流 1.5 A, 光电倍增管电压 -1 200 V, 增益 32, 分析波长范围 476~766 nm, 扫描步距 1 nm, 采样速率 300 nm·min⁻¹。每个样本扫描 30 次, 保存 5 条光谱曲线, 以其平均光谱作为该样本的原始光谱。

1.3 高斯拟合

光谱分析中峰位较接近时会不可避免地存在光谱重叠干扰, 这对从光谱中提取微弱化学成分变化信息会造成比较大的困难。利用高斯函数对光谱曲线进行拟合解谱, 不仅可以表达谱图中峰形、峰高和峰位等具有明确物理意义的参数, 也可以在保持光谱有效信息的基础上, 将实际测量的原始光谱数据统一为少数几个具有客观评价意义的确定高斯特征参数, 简化光谱曲线的数据矩阵, 实现数据的压缩, 从而建立合适的光谱数据定量分析模型。

谱图分离处理通常以最小方差准则的曲线拟合法进行计算而获得。设实际采样的数据为 (x_i, y_i) , 高斯拟合函数记为 $f(x, P)$, 其中 P 是以各拟合曲线参数为元素构成的矢量。调整 P 值, 使组合函数的波形和实测输出波形的差值二乘方和

$$E(P) = \sum_{i=1}^n Z_i \{f(x_i, P) - y_i\}^2 \quad (1)$$

最小, 以此来确定 P 而达到解谱的目的。 $E(P)$ 称为目标函数, Z_i 为加权系数, 引入 Z_i 的目的是为了进行高精度的拟合。 $f(x, P)$ 高斯函数系写作

$$f(x, P) = \sum_{i=1}^n A_i \times h_i \exp\left\{-\frac{(x - \lambda_i)^2}{2\omega_i^2}\right\} + B \quad (2)$$

式(1)中待估参数 h_i , λ_i 和 ω_i , 分别代表函数系中各高斯曲线的峰高、峰位和半峰宽度, A_i 提供各高斯曲线在函数系中所占比重信息, B 表示光谱曲线的基底。

1.4 遗传算法

遗传算法作为模拟生物进化机制的随机优化算法, 对非线性式(1)进行最小值优化计算时, 克服了传统优化算法需

计算目标函数的梯度造成计算量大, 曲线拟合中样品元素及谱线未知、背景基线不确定、模型初始参数不准确等一系列困难, 而是以适应度函数为标准, 通过对群体中的个体施加选择、交叉、变异等遗传操作建立起迭代过程, 在复杂、多峰值、不可微的大矢量空间中迅速有效地寻找到全局最优解, 鲁棒性强, 谱峰拟合能力好。

本研究的适应度函数采用测量与拟合光谱误差的倒数形式; 为了平衡全局搜索能力和收敛速度, 选择每隔 10 代用新产生的个体替换种群中最差个体的进化策略, 以保持种群的多样性。

1.5 BP 神经网络模型

建立了一个三层的人工神经网络结构, 各层传递函数采用 S 型(Sigmoid)函数。目标误差为 0.000 1, 网络指定参数中学习速率为 0.05, 设定训练迭代次数为 50 000 次。

2 结果与分析

2.1 真假发票的光致发光光谱特征

真假发票的典型光致发光光谱曲线如图 2 所示。每种发票随意选取 10 条光谱。图中横坐标为波长, 范围是 476~766 nm, 纵坐标为光谱强度。从图 2 中可以看出, 真假发票的激光光致发光光谱在 566~669 nm 波段都存在一个明显的宽峰, 但由于荧光峰的重叠范围较大, 仅从峰值位置上尚不能将其区分开来, 必须要结合重叠谱的解析方法对光谱数据进行处理, 从而建立起发票真伪鉴别模型。

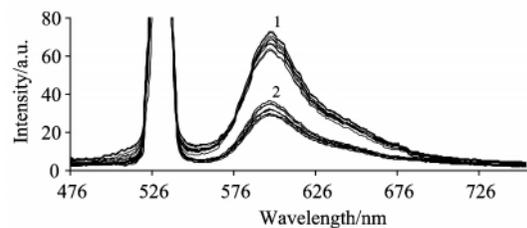


Fig 2 Laser-induced photoluminescence spectra of authentic and false invoice samples

1: False invoice; 2: Authentic invoice

2.2 基于遗传算法的高斯拟合

基于遗传算法对真假发票的 80 个样本进行高斯拟合, 提取特征参数。表 1 为拟合得到的各参数的平均值, 可以看出真假发票的光谱主峰位置的平均值相差 0.8 nm, 所占比重明显不同, 为发票的真伪鉴别提供了较精确的结果。图 3 和图 4 分别是对真假发票的拟合效果图, 真发票的确定系数 $R^2=0.997 89$, 相对标准差 $RSD=0.017 052$; 假发票的确定系数 $R^2=0.996 83$, 相对标准差 $RSD=0.022 362$ 。表明对光谱数据进行高斯拟合后, 原始光谱的有效信息丢失量少, 提取的 13 个特征参量约只为 566~669 nm 波段共 104 个点的 12.5%, 实现了将原来的光谱信息统一为少数唯一确定的高斯参量, 有效地进行了数据压缩。为达到更高的鉴别精度, 实验采用 BP 神经网络结合基于遗传算法的高斯拟合法建立发票真伪鉴别定量分析模型。

Table 1 Fitting results of three Gaussian peaks of authentic and false invoice spectra

特征参量	假发票			真发票		
	高斯峰 1	高斯峰 2	高斯峰 3	高斯峰 1	高斯峰 2	高斯峰 3
A	0.86	0.89	0.64	0.95	0.91	0.58
h	48.72	42.12	31.29	25.29	21.13	13.49
λ/nm	595.1	605.2	624.3	594.3	605.1	624.2
ω/nm	11.81	9.22	17.54	11.52	9.29	19.02
B	12.14			6.03		

2.3 BP 神经网络预测模型

由于光谱数据是庞大的低信息密度高维空间数据, 直接将全部光谱数据(476~766 nm)用来建模, 容易导致输入变量太多模型难以收敛, 增加建模难度, 同时无关的噪声信息也会融入到模型中反而降低模型的预测精度。为此将基于遗传算法拟合计算提取的 13 个特征参数和拟合效果评估参数 R^2 和 RSD 作为神经网络的输入建立预测模型。

从 566~669 nm 有效光谱范围中取每个试验样本的 13 个特征参数和 2 个评估参数, 分别挑选不同个数作为神经网络的输入变量, 发票真伪作为神经网络的输出变量(真发票设为 1, 假发票设为 2), 通过调整隐含层的节点数来优化网络结构, 建立 3 种不同的 BP 神经网络模型。真假发票 60 个建模样本, 对 20 个未知样本的真伪进行鉴别。设定预测结果偏差在 ± 0.1 内为预测正确。表 2 显示了不同输入变量的预测结果。从表 2 可以看到 3 种模型的假发票被识别能力较理想, 建模和预测识别准确率在 80% 以上。相比模型 1, 模型 2 的输入变量进一步全面描述了原始光谱的精细结构, 建模样

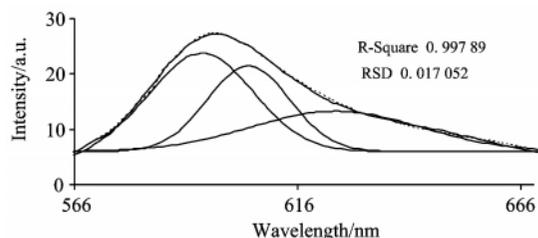


Fig 3 Comparison between fitting curves of authentic invoice and experimental spectrum

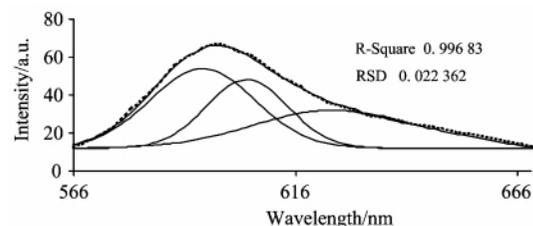


Fig 4 Comparison between fitting curves of false invoice and experimental spectrum

本和预测样本的识别准确率显著增高。而模型 3 的输入变量所包含的 2 个评估参数, 还描述了高斯拟合曲线与原始光谱曲线之间的相关程度信息, 相比模型 2 更充分解释了真假发票原光谱的有效信息, 因此模型质量最高, 对建模样本和预测样本的识别准确率均达到 100%, 可以作为鉴别发票真伪的有效手段。由此可见, 用激光光致发光光谱技术结合基于遗传算法的高斯拟合和 BP 神经网络法建立发票真伪鉴别模型是可行的。

Table 2 Effect of different inputs for BP model

输入变量个数	最佳网络结构	建模				预测			
		识别准确率/%	真发票识别准确率/%	假发票识别准确率/%	误判个数	识别准确率/%	真发票识别准确率/%	假发票识别准确率/%	误判个数
模型 1: 9($h_1, \lambda_1, \omega_1, h_2, \lambda_2, \omega_2, h_3, \lambda_3, \omega_3$)	9-13-1	86.67	83.33	90.00	8	75.00	70.00	80.00	5
模型 2: 13($A_1, h_1, \lambda_1, \omega_1, A_2, h_2, \lambda_2, \omega_2, A_3, h_3, \lambda_3, \omega_3, B$)	13-17-1	98.33	96.67	100.00	1	100.00	100.00	100.00	0
模型 3: 15($A_1, h_1, \lambda_1, \omega_1, A_2, h_2, \lambda_2, \omega_2, A_3, h_3, \lambda_3, \omega_3, B, R^2, \text{RSD}$)	15-20-1	100.00	100.00	100.00	0	100.00	100.00	100.00	0

3 结论

本研究利用激光光致发光光谱分析技术对发票真伪鉴别作了初步探讨, 结合基于遗传算法的高斯拟合和 BP 神经网络法建立发票真伪鉴别模型。应用 532 nm 激发光照射普通真假发票的监制章产生激光光致发光光谱, 通过高斯拟合对

谱线重叠严重的 566~669 nm 波段光谱特征进行有效的提取和简化。经比较分析, 将压缩后 13 个特征参数和 2 个评估参数作为神经网络的输入, 提高了模型的精度, 对建模样本和预测样本的识别准确率均达到 100%, 实现了运用激光光致发光光谱技术快速、准确地鉴别发票的真伪, 为发票的快速鉴别提供了一种新方法。

References

- [1] ZHAO Qi-zhu(赵启柱). Journal of Huainan Institute of Technology(淮南工业学院学报), 2002, 22(2): 30.
- [2] CHEN Shu-gen, JIANG He-ping, ZHOU Gao-bei(陈树根, 江和平, 周高杯). Infrared and Laser Engineering(红外与激光工程), 2001, 30(4): 174.
- [3] Ohta H, Portugall O, Ubrig N, et al. Journal of Low Temperature Physics, 2010, 159(1-2): 203.
- [4] Yan Bing, Lin Lixia, Wu Jianhua, et al. Journal of Fluorescence, 2011, 21(1): 203.
- [5] Holland J H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, Ann Arbor, MI, 1975.
- [6] Taysi N, Göğüş M T, Özakça M. Computational Optimization and Applications, 2008, 41(3): 377.
- [7] Sabri Kocer, M Rahmi Canal. Journal of Medical Systems, 2009. Available Online. <http://springer.lib.tsinghua.edu.cn/content/j362366625471648/fulltext.pdf>.

Identification of Invoice Based on Laser-Induced Photoluminescence Spectrum

YANG Qin¹, YANG Yong², TIAN Yong-hong¹

1. School of Physical Science and Technology, Yangtze University, Jingzhou 434023, China

2. School of Mathematics and Physics, China University of Geosciences, Wuhan 430074, China

Abstract The rapid identification of invoice authenticity was studied based on laser-induced photoluminescence spectrum. First, the spectral curves of eighty invoice samples were obtained by laser-induced photoluminescence detection system, and genetic algorithm (GA) was applied to fit and separate overlapped spectral region between 566 and 669 nm by three Gaussian peaks. Spectral feature parameters extracted by GA were employed as the inputs of BP neural networks, and then an identification model was built. One hundred and four data were converted to 13 Gaussian parameters, and for authentic and false invoices the coefficients of determination (R^2) were 0.997 89 and 0.996 83 and the relative standard deviations (RSD) were 0.017 052 and 0.022 362, respectively. It was showed that Gaussian fitting algorithm could not only simplify the parameters of models, but also improve the explanation of analysis models. Through comparison analysis of the results, it was found that the model, whose thirteen feature parameters and two evaluated parameters were all applied as BP inputs, was the best, and the corrected identification rates of sixty calibration samples and twenty validation samples were both 100%. So the identification method studied in the present research played a good role in the classification and identification, and offered a new approach to the rapid identification of invoice authenticity.

Keywords Laser-induced photoluminescence; Invoice; Genetic algorithm; Gaussian fitting; BP neural network; Identification

(Received Mar. 10, 2011; accepted Jul. 16, 2011)