第37卷，第12期
2017年12月

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

Vol. 37, No. 12, pp3828-3832
December, 2017

支持向量机的动物血液光谱特征提取和识别分类

卢鹏飞[1]，范 雅[1]，周林华[1*]，钱 军[2]，刘林娜[2]，赵思言[2]，孔之丰[3]，高 斌[1]

1. 长春理工大学理学院，吉林 长春 130022
2. 中国农业科学院长春兽医研究所，吉林 长春 130122
3. 西安交通大学数学与统计学院，陕西 西安 710048

摘 要 利用光谱检测和数据挖掘实现不同种类动物血液光谱数据的精确识别与分类具有重要意义，目前尚未见到较为完善及普适的相关研究报道。实验采集了鸽、鸡、鼠、羊四种动物全血和红细胞溶液(浓度为1%)的荧光光谱数据；基于小波变换的软阈值去噪方法，首先对原始光谱数据进行去噪处理，并确定了717个原始特征(包括荧光峰强度值、荧光峰连线斜率等4类特征)；提出以“区分度统计量”为核心的特征提取方法，结合主成分分析法和平均影响值算法，实现了对717个原始特征到2个识别特征的高效筛选；进一步建立了径向基核函数的支持向量机分类器，对四类不同动物的全血荧光光谱数据实现了准确率为100%的识别分类，对红细胞荧光光谱数据实现了94.69%～99.12%的识别率；最后蒙特卡洛交叉验证的结果表明所提出的思路和方法对于动物全血溶液的识别分类具有较好的泛化能力，能对荧光光谱数据进行准确的识别分类，因此能够在进出口检查、食品安全、医药等领域发挥重要作用。针对动物血液荧光光谱，提出的基于“区分度统计量”的特征提取方法，相比于传统的人为特征选取方法，能够从大量原始特征中自动提取少量且有效的识别特征，具有较强的普适性和高效性，为其他领域的光谱特征提取和识别分类提供了一种新的思路。

关键词 动物血液；荧光光谱；识别分类；特征提取；支持向量机

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2017)12-3828-05

引 言

动物血液鉴定在进出口、食品安全与卫生等领域有着众多实际需求。在进出口领域，需要血液鉴别以避免非法进出口；在食品安全与卫生领域，需要高效血液鉴别方法来提高生产和监管效率，等等。因此，开发动物血液快速、准确的识别分类技术具有重要应用价值。

近些年来，光谱技术正逐渐成为生物研究领域的有力工具。例如：冯尚源等[1]利用人体血浆的表面增强拉曼光谱结合多变量统计方法对胃癌的无损诊断分析进行了研究；王畅等利用正常奶牛和患病奶牛(患乳腺炎)血液样品的荧光光谱在某些区域的变化鉴别了奶牛的血液样品是否异常；褚璇等[2]采用主成分支持向量机判别模型对掺假山茶油和未掺假山茶油在1 000～2 500 nm波长范围内的吸收光谱进行了识别鉴定；马伟等使用簇类的独立软模式分类法对健康万寿菊和患黑斑病万寿菊的近红外光谱进行了识别。以上均是对人体血液及同一物种动物的血液、植物的生物属性鉴定方法的研究。

目前为止，利用血液光谱进行不同动物种类鉴定分类的研究结果较少。2016年，万雄等[3]发现可以通过犬、猫、鸡三类动物全血样品近红外透射光谱相关系数的差异来进行全血鉴定判别，但受实验设备和技术的影响较大。2016年，白鹏利等[4]以三种不同的动物血样以及21个人血液作为分析对象，采用主成分分析结合拉曼光谱进行血液定性识别检测，但其方法过于依赖人的判断，未能实现计算机程序自动识别。以上研究，为进一步利用血液光谱开发更高效的动物血液识别分类技术提供了思路。

荧光光谱能够探测到生物体中分子的化学组成结构及分子与分子、分子与周围环境相互作用的信息，从而通过研究荧光光谱可以获得生物体细胞的组成和代谢信息。我们利用荧光分光光度计得到了鸽、鸡、鼠、羊四种动物的全血荧光

收稿日期：2017-01-08，修订日期：2017-05-14

基金项目：国家自然科学基金项目(1120420，11426045)资助

作者简介：卢鹏飞，1991年生，长春理工大学理学院硕士研究生 e-mail：921010752@qq.com

*通讯联系人 e-mail：zhoulh@cust.edu.cn

光谱数据和红细胞荧光光谱数据。基于提出的“区分度统计量法”，实现了动物血液溶液荧光光谱特征的自动提取；进一步，利用径向基核函数支持向量机分类器实现了不同动物血液溶液的准确分类。

1　实验部分

1.1　样品及仪器

实验采集了鸽、鸡、鼠、羊四种动物的全血和红细胞样本，将抗凝剂(柠檬酸钠)预先加入到采血容器中，待采集血液后使用 0.85% NaCl 溶液(生理盐水)稀释至 1%浓度全血溶液待用；同样采集血液后先洗涤为血红细胞溶液，再使用 0.85% NaCl 溶液(生理盐水)稀释至 1%浓度血红细胞溶液待用。使用 Cary Eclipse 荧光分光光度计于室温下扫描待用的全血溶液和红细胞溶液样本，采集发射光谱数据。实验的具体参数和设置为：采样类型(荧光)、扫描方式(发射)、激光波长(200 nm)、激光狭缝(5 nm)、发射狭缝(10 nm)、扫描速度(600 nm·min^{-1})、数据间隔(1 nm)、平均时间(0.1 s)。

图 1 是对两类样本溶液扫描得到的荧光光谱数据曲线。通过对比发现无论是全血溶液或是红细胞溶液，随着波长的增加，不同动物血液样本的光谱强度变化趋势较为一致，难以直接区分。同时，所有光谱曲线均有六个较为明显的荧光峰，分别在 320，370，500，610，750 和 830 nm 附近。

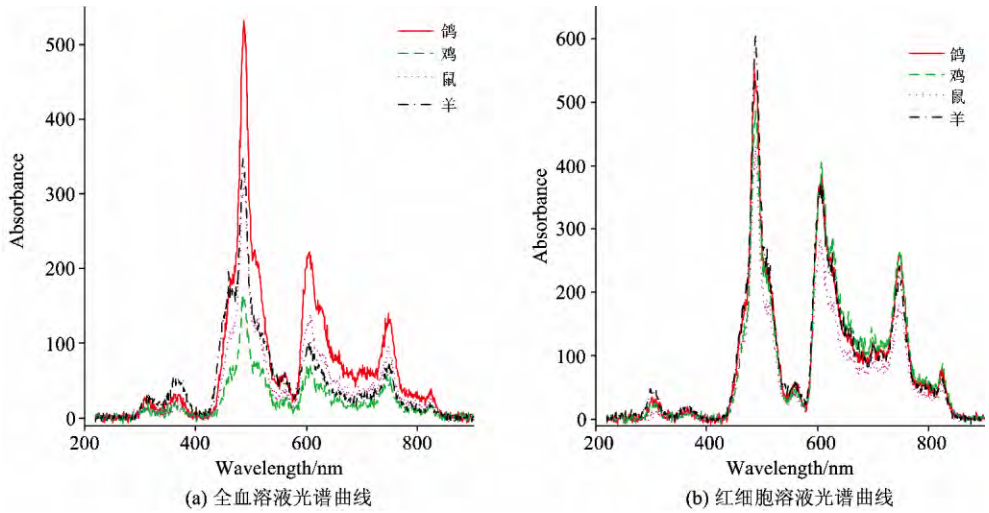

图 1　样本光谱图

Fig. 1　Spectra of whole blood solution (a) and red blood cell solution (b)

1.2　方法

整体思路和具体方法包括：数据预处理、确定原始特征、特征筛选、特征提取、支持向量机(support vector machine，SVM)分类器建立和参数寻优、识别分类和交叉验证评价，如图 2 所示。

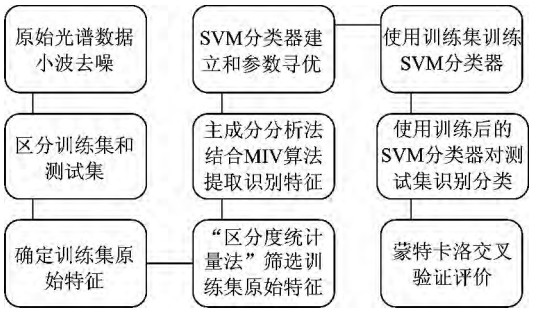

图 2　整体思路流程图

Fig. 2　Flow chart of principle idea

1.2.1　“区分度统计量法”筛选特征

遵循“类内稳定、类间离散”的特征选取原则，提出“区分度统计量法”筛选原始特征。具体步骤如下：

(1)原始特征值记为

x_{ijk}，($i=1,2,\cdots,m$；$j=1,2,\cdots,p$；$k=1,2,\cdots,q$)

其中 i 为原始特征，j 为动物类别，k 为同一动物类别的不同样本。规范化处理得到特征转换值

$$y_{ijk}=\frac{x_{ijk}-\min\limits_{1\leqslant k\leqslant q}x_{ijk}}{\max\limits_{1\leqslant k\leqslant q}x_{ijk}-\min\limits_{1\leqslant k\leqslant q}x_{ijk}}$$

(2)将特征转换值 y_{ijk} 关于样本 $k=1,2,\cdots,q$ 计算平均值，并关于动物类别排序：$\bar{y}_i^{(1)}\leqslant\bar{y}_i^{(2)}\leqslant\cdots\leqslant\bar{y}_i^{(p)}$；同时，特征转换值 y_{ijk} 关于第 i 个原始特征、第 j 个动物类别的样本方差记为 $S_i(j)$。

定义第 i 个原始特征的区分度统计量为

$$FS(i)=\left(1-\sqrt{\sum_{j=2}^{p}|\bar{y}_i^{(j)}-\bar{y}_i^{(j-1)}|^2}\right)\sqrt{\sum_{j=1}^{p}S_i(j)}$$

其中第一个根号刻画不同动物类别间的特征转换值集散程度；第二个根号为同一动物类别特征转换值集中程度的总体量化值。显然，区分度统计量越小表明该原始特征的区分度越高。因此，借助于区分度统计量，可以量化评价原始特征对分类的有效性，然后筛选出区分能力较强的原始特征。

1.2.2　径向基核函数 SVM 分类器

支持向量机是在统计学习理论和结构风险最小原理基础上发展起来的一种分类识别方法[5]，我们选择核函数支持向

量机作为识别分类器。研究表明，当缺少过程的先验知识时，相比较其他核函数，径向基核函数参数少，而且具有更好的性能[6]。因此，选择径向基核函数进行 SVM 分类识别，径向基核函数定义如下

$$K(x, y) = \exp(-\|x-y\|^2/2\sigma^2)$$

径向基核函数支持向量机的惩罚参数 C 和核函数参数 σ^2 会直接影响到 SVM 的分类效果，使用了由 Kennedy 和 Eberhart 提出的粒子群优化算法[7]进行参数寻优。

1.2.3 蒙特卡洛交叉验证

交叉验证可以评价分类器识别性能的准确性和稳定性。使用基于蒙特卡洛抽样思想的 K-fold cross-validation 对识别分类效果进行交叉验证。其具体步骤为：将原始数据随机均分成 K 组，将每组样本数据分别做一次验证集，其余的 $K-1$ 组子集样本数据作为训练集，该次验证会进行 K 次识别；重复上述过程 n 次(取 100 次)，最终综合 $n\times K$ 次测试计算识别正确率的均值和标准差。

3 结果与讨论

3.1 基于小波的数据预处理

采用基于小波变换的软阈值去噪方法[10]。以“db3”作为小波母函数，选择”Heursure”阈值选取规则和“sln”阈值调整方法，对原始光谱进行去噪处理。图 3(b)—(f)分别显示了小波分解层数为 1 至 5 层时全血溶液荧光光谱的去噪结果。从图中可以看出，当层数为 1 时，光谱去噪效果不明显；当层数为 4 和 5 时，部分荧光峰由于被过度光滑处理而被去除了。与层数 2 相比，当层数为 3 时原始光谱不仅能被更好的去除噪声，还能保留原始光谱中有助于识别的特征。因此，本文选取 $N=3$ 对原始数据进行去噪。

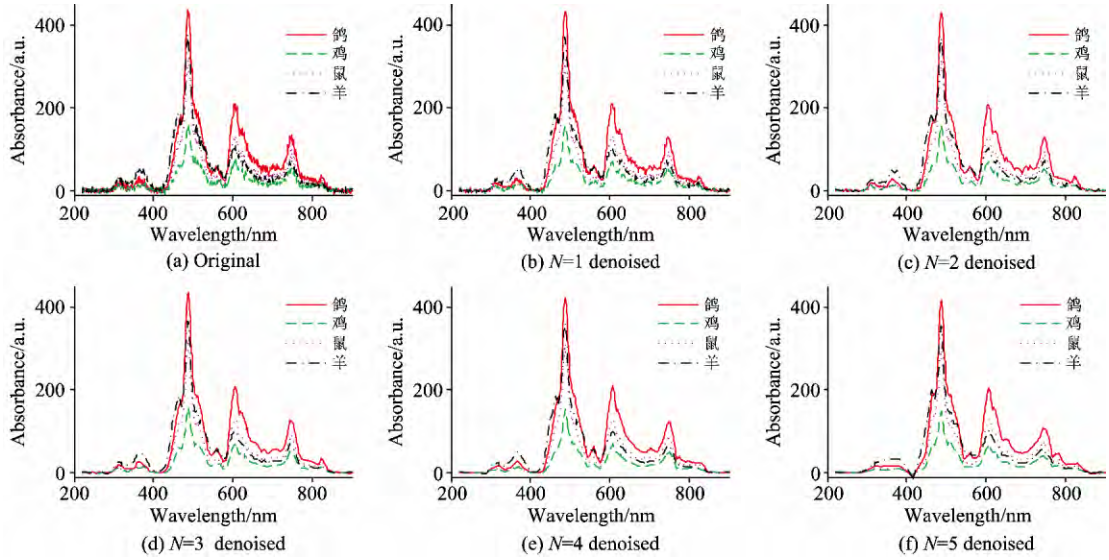

图 3 不同分解层数时的去噪效果图

Fig. 3 Denoising results of different layers

3.2 原始特征的确定

在进行特征筛选和提取之前，需要确定原始特征。在确定原始特征时，不仅考虑了每个波长处的光谱强度，还考虑了荧光峰强度值、荧光峰连线斜率、荧光峰差值等特征。原始光谱共有 681 个光谱强度值和 6 个明显荧光峰，因此使用的原始特征如表 1 所示。

表 1 原始特征的种类和个数

Table 1 Types and number of original features

原始特征类别	个数
光谱强度值	681
荧光峰强度值	6
荧光峰连线斜率	15
荧光峰差值	15

3.3 基于区分度统计量的识别特征确定

首先，基于“区分度统计量法”计算原始特征的区分度统计值，进行从小到大排序，筛选出前 100 个原始特征；然后，再使用主成分分析法选取贡献率最大的前 15 个主成分作为初步选定变量；最后，使用平均影响值(mean impact value，MIV)算法[9]对训练集数据处理，筛选出对支持向量机输出影响明显较大的主成分。

对于全血溶液光谱数据，通过计算得到不同主成分对模型输出的影响值及其排序。结果发现前两个主成分的 MIV 绝对值远大于其他主成分，累计贡献率高达 99.58%。因此选取前两个主成分作为支持向量机的输入特征，具体结果如表 2 所示。图 4 是全血溶液光谱识别特征数据的散点图，图中可见不同种类动物的输入特征分别呈聚集状，类别间区分明显。

3.4 SVM 识别结果

本次识别借助 MATLAB 中的支持向量机工具箱进行识别计算。设置惩罚参数 C 和核函数参数 σ^2 的初始设定值分别为 2 和 1。通过 MATLAB 编程，建立了使用 6 个子分类器

表 2　全血溶液的主成分 MIV 值

Table 2　MIV value of the principal components of whole blood 1%

主成分序号	贡献率/%	MIV 值	绝对值排序
1	90.007 1	0.024 3	1
2	9.582 3	−0.024 3	2

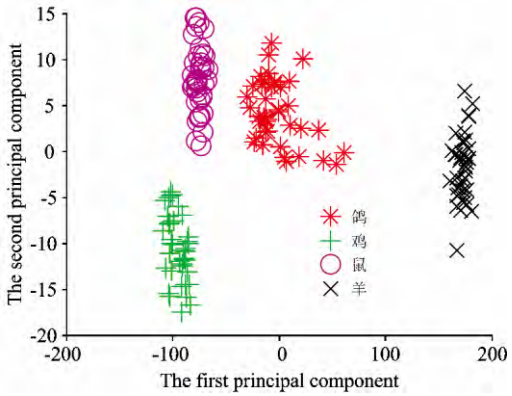

图 4　全血溶液识别特征散点图

Fig. 4　Scatter plots of the recognition features of whole blood 1%

投票的多类别识别分类器。然后，使用粒子群算法优化迭代，寻找使得训练集分类准确率最高的$\{C, \sigma^2\}$参数组合。设置初始种群数量为 20，最大进化代数 200，c_1 和 c_2 分别初始化为 1.5 和 1.7，所需优化参数$\{C, \sigma^2\}$的搜索范围分别设定为$[0.1, 1\,000]$，$[0.01, 1\,000]$，通过计算得到最优参数组合为$\{1.628, 2\}$。

使用已确定的输入特征数据进行标准化处理后作为训练样本，然后使用参数优化后的 SVM 和 160 个训练样本进行模型训练，最后将 40 个测试样本的输入特征输入支持向量机进行识别。识别结果如图 5 所示，全血溶液测试集的识别正确率达到了 100%，获得了满意的识别效果。

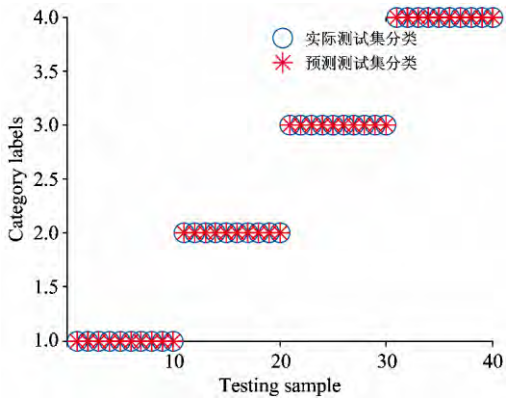

图 5　全血溶液测试集的识别结果

Fig. 5　Recognition results of test set of whole blood

3.5　蒙特卡洛交叉验证结果

分别取 $K=2, 4, 6, 8, 10$ 对全血溶液样本进行 100 次随机抽样识别，使用蒙特卡洛交叉验证对识别体系进行验证。验证结果如表 3 所示，结果表明在不同的分组交叉验证情形下，全血溶液测试集的识别分类准确率均为 100%。表明提出的特征提取方法和设计的分类器对于浓度为 1%的全血溶液而言，受实验影响较小，具有较高的独立于数据集的泛化能力。

表 3　全血溶液的交叉验证结果

Table 3　Results of cross validation of whole blood

K 取值	平均准确率/%	标准偏差
2	100	0
4	100	0
6	100	0
8	100	0
10	100	0

3.6　红细胞溶液识别结果分析

对于浓度为 1%的红细胞溶液，使用处理全血溶液一致的方法进行了训练和识别，其蒙特卡洛交叉验证结果如表 4 所示。当 K 为 2 时，红细胞溶液的 100 次随机抽样的识别准确率最高，平均准确率达到 99.12%，标准差为 0.742 4；当 K 为 4，6，8，10 时，红细胞溶液的识别准确率和稳定性相对较差，而且存在过拟合。因此，应当考虑使用全血溶液的荧光光谱数据对不同动物血液进行识别分类。

表 4　红细胞溶液的交叉验证结果

Table 4　Results of cross validation of red blood cell

K 取值	平均准确率/%	标准偏差
2	99.12	0.742 4
4	97.423 1	2.012 7
6	94.861 1	2.744 7
8	94.642 9	2.717 4
10	94.25	4.666 5

4　结　论

以识别不同动物血液溶液为目标，提出了基于“区分度统计量法”的特征提取方法，并建立了径向基核函数支持向量机分类器，最终对全血溶液荧光光谱数据实现了准确率为 100%的识别分类。蒙特卡洛交叉验证的结果也表明提出的思路和方法具有较好的泛化能力，因此能对同类条件下获得的荧光光谱数据进行准确识别分类。

基于“区分度统计量”法的特征提取技术与人工提取特征的方法相比，不仅节省人力，还能更加准确的提取出有利于分类的特征。在仅仅使用了两个识别特征的情况下，达到了准确且稳定的识别效果，提高了识别运算效率。

由于条件所限，目前只对四种动物的血液溶液荧光光谱数据进行识别分类。期望能在后继工作中，对更多动物种类的血液溶液进行光谱数据识别分类研究，建立动物血液光谱数据资料库，为进出口、医学、食品安全等领域提供快速、准确且节省人力的动物血液分类检测与鉴别方法。

References

[1] Feng S Y, Pan J J, Wu Y A, et al. Science China Life Sciences, 2011, 54(9): 828.

[2] WANG Chang, WANG Le-xin, ZHAO Zhi-min(**王 畅, 王乐新, 赵志敏**). Applied Laser(**应用激光**), 2013, 33(4): 456.

[3] WAN Xiong, WANG Jian, LIU Peng-xi, et al(**万 雄, 王 建, 刘鹏稀, 等**). Spectroscopy and Spectral Analysis(**光谱学与光谱分析**), 2016, 36(1): 80.

[4] BAI Peng-li, WANG Jun, YIN Huan-cai, et al(**白鹏利, 王 钧, 尹焕才, 等**). The Journal of Light Scattering(**光散射学报**), 2016, 28(2): 163.

[5] DING Shi-fei, QI Bing-juan, TAN Hong-yan(**丁世飞, 齐炳娟, 谭红艳**). Journal of University of Electronic Science and Technology of China(**电子科技大学学报**), 2011, 40(1): 2.

[6] CHEN Sheng-bing, WANG Xiao-feng(**陈圣兵, 王晓峰**). Computer Engineering and Applications(**计算机工程与应用**), 2012, 48(7): 20.

[7] GU Wen-cheng, CHAI Bao-ren, TENG Yan-ping(**谷文成, 柴宝仁, 滕艳平**). Transactions of Beijing Institute of Technology(**北京理工大学学报**), 2014, 34(7): 705.

[8] XIA Shu-hua(**夏淑华**). Computer Simulation(**计算机仿真**), 2011, 28(4): 332.

[9] NIE Ming, ZHOU Ji-heng, YANG Rong-sheng, et al(**聂 铭, 周冀衡, 杨荣生, 等**). Acta Tabacaria Sinica(**中国烟草学报**), 2014, 20(6): 50.

Feature Extraction and Classification of Animal Blood Spectra with Support Vector Machine

LU Peng-fei[1], FAN Ya[1], ZHOU Lin-hua[1]*, QIAN Jun[2], LIU Lin-na[2], ZHAO Si-yan[2], KONG Zhi-feng[3], GAO Bin[1]

1. School of Science, Changchun University of Science and Technology, Changchun 130022, China
2. Changchun Veterinary Institute, Chinese Academic Agricultural Sciences, Changchun 130122, China
3. School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710048, China

Abstract It is of great significance to study how to use spectral detection technology and data mining technology to realize the accurate identification and classification of different animal blood spectral data, and it has not yet seen relevant complete research conclusions and methods on animal blood identification and classification. Therefore, the authors collected fluorescence spectra data of four kinds of animals, including pigeon, chicken, mouse and sheep. Based on the soft threshold denoising method of wavelet transform, the original spectral data were denoised, and the 717 original features were determined. Following the approach of "Distinguish statistic" proposed by the authors, 717 original features were extracted into 2 finally input features. Based on support vector machine, the whole blood solution of different animals were 100% recognized, while the red cell blood solution of different animals were 94.69%～99.12% correctly recognized. Finally, the Monte Carlo cross validation revealed that the method used in this paperhad a great generalization ability for whole blood solution of different animals, which can play an important role in the import and export inspection, food safety, medicine and other fields.

Keywords Animal blood; Fluorescence spectrum; Classification; Feature extraction; Support vector machine

(Received Jan. 8, 2017; accepted May 14, 2017)

* Corresponding author