文章编号: 1000-5641(2017)02-0069-06

基于维度约束的距离测度学习算法

房 娟, 刘洪英, 李庆利

(华东师范大学 信息与科学技术学院 上海市多维度信息处理重点实验室, 上海 200241)

摘要: 为提高分类精度, 通过距离测度学习可以得到样本在新的特征空间里新的表示. 针对马氏距离未考虑不同类别样本维度间相关性存在差异这一缺陷, 提出了一种新的有监督的距离测度学习算法, 即独立-差别分量分析方法(Independent Discrimi-Native Component Analysis, I-DCA), 并将其运用于基于k近邻分类器的运动神经与感觉神经分类中. 作为对照, 还详细分析了已有的相关分量分析方法(Relevant Component Analysis, RCA)和差别分量分析方法(Discrimi-Native Component Analysis, DCA)这两种距离测度学习算法. 实验结果表明, 改进算法的分类精度相较于马氏距离提高了近 45%, 相较于RCA与DCA也提高了 15%左右, 分类精度的提高说明了改进算法的有效性.

关键词: 距离测度学习; 有监督学习; k近邻分类; 显微高光谱; 神经分类

中图分类号: TP181 **文献标志码**: A **DOI:** 10.3969/j.issn.1000-5641.2017.02.009

Learning distance metrics with dimension constraints

FANG Juan, LIU Hong-ying, LI Qing-li

(*School of Information Technology, Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai* 200241, *China*)

Abstract: In order to improve the classification accuracy, the new representation of samples can be gotten by distance metric learning. According to mahalanobis distance does not take the difference of the relativity between different classes of sample dimensions into consideration. A new supervised distance metric learning algorithm called independent discrimi-native component analysis(I-DCA) is proposed and applied to classify the motor and sensory nerve based on k nearest neighbor (kNN) algorithm. By contrast, the article also involves the analysis of two existing distance metric learning algorithms in detail, the relevant component analysis (RCA) and the discrimi-native component analysis(DCA). Compared with the mahalanobis distance, the results indicate that the classification precision of the improved algorithm increases by nearly 45%, and it is also greater than 15% compared to the RCA and DCA method. The improved classification precision shows the effectiveness of the new algorithm applied in nerve classification.

Key words: distance metric learning; supervised learning; kNN classification; microscopic

收稿日期: 2015-07-08

基金项目: 国家自然科学基金(61240006)

第一作者: 房 娟, 女, 硕士研究生, 研究方向为多维信息处理. E-mail: fjoanna@126.com.

通信作者: 刘洪英, 女, 副教授, 研究方向为多维信息处理. E-mail: hyliu@ee.ecnu.edu.cn.

hyperspectral imaging; neural classification

0 引 言

寻求一种快速、简便而准确的方法来鉴别神经束的功能成分一直是显微外科医生们面临的重大课题, 因为在修复损伤的周围神经时, 只有将相匹配的两断端神经纤维(主要包含感觉神经以及运动神经)进行正确的镜面吻合, 才能保证神经修复的效果[1]. 现有的神经束性质识别方法包含解剖学方法、电生理法、组织化学法、同位素法等, 鉴于高光谱成像技术在生物医学方面的应用越来越广泛, 已有研究将该技术应用于周围神经的分类中, 研究[1-3]表明两种神经轴突及髓鞘部分的光谱曲线存在差异. 基于k近邻分类器, 本文将进一步对神经分类的可行性进行有效的探究.

k值的选择、距离的度量函数和分类决策规则是影响k近邻分类器性能的3个基本要素, 而一个合适的距离测度是衡量样本点间相似度的关键. 距离测度学习旨在将样本从原始空间变换到新的特征空间, 使得新的样本表示具有更合适的距离测度以用于分类[4]. 目前已有很多算法被相继提出, 这些算法主要分为有监督和无监督两类. 有监督的距离测度学习算法利用同类样本之间形成的对等约束和异类样本之间形成的不对等约束, 主要算法有RCA[5]、近邻分量分析方法[6](Neighborhood Component Analysis, NCA)、基于凸优化的全局距离测度学习方法[7](Probablistic Global Distance Metric Learning, PGDM)、DCA[8] 等; 无监督的算法无类别信息参与, 多用于基于谱分析的降维算法中, 主要算法有主成分分析(Principle Component Analysis, PCA)、局部保持投影(Locality Preserving Projection, LPP)、局部线性嵌入(Locally Linear Embedding, LLE)等.

通过学习研究各种已有的算法, 分析算法原理, 并针对马氏距离只考虑总体样本各维度相似性, 而未考虑不同类别样本维度间相关性存在差异这一问题, 本文提出一种新的有监督的距离测度学习算法即I-DCA, 并基于k近邻分类器用该算法对兔子运动神经与感觉神经的显微高光谱图像数据进行分类. 作为对比, 同时也计算了相同条件下运用欧氏距离、马氏距离、RCA和DCA方法的分类结果.

1 算法原理

1.1 相关基础

假设一数据集$X=\{x_1,x_2,x_3,...,x_n|x_i\in\mathbf{R}^{1\times m}\}$, 其中, n表示样本数目, m表示样本的维数, 数据集各样本的标签信息为$L=\{l_1,l_2,l_3,...,l_n\}$, 表示各样本所属的类别. 距离测度学习的目的是基于数据集利用或者不利用标签信息学习得到距离测度矩阵M(M为半正定矩阵), 使得样本x_i和x_j之间的距离度量表示为$d_M(x_i,x_j)=\sqrt{(x_i-x_j)^{\mathrm{T}}M(x_i-x_j)}$, 当$M$为单位阵时, d表示样本间的欧氏距离; 若将M分解为$M=A^{\mathrm{T}}A$, 则A可作为变换矩阵对样本进行线性或非线性变换, 使其成为另一种更具类别区分性的表示形式[9-11], 即

$$d_M(x_i,x_j)=d_A(x_i,x_j)=\sqrt{(Ax_i-Ax_j)^{\mathrm{T}}(Ax_i-Ax_j)}. \tag{1}$$

相对于欧式距离直接指示空间中两样本点间的实际距离, 印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出了另一种计算两个未知样本集相似度的方法, 即马氏距离. 马氏距离考虑到数据集各种特性之间存在联系, 利用协方差矩阵$\sum$来反映样本各维度信息之间

的相关性, 经白化变换后矩阵 $\sum^{-1/2}$ 可以重新分配数据集各维度的权重使各维度相互独立. 空间变换示意图如图1所示, 呈椭球分布的样本在经过一系列变换后在新的空间呈球状分布; 而原始空间内的马氏距离等价于在呈球状分布空间里所求的欧氏距离.

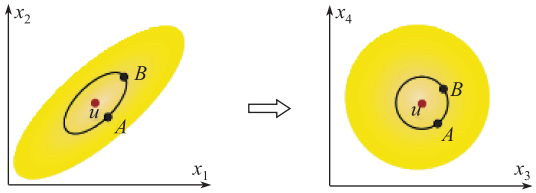

图 1　空间变换示意图

Fig. 1　Schematic illustration of the space transformation

1.2　距离测度学习算法

对于同一数据集下不同类别的样本集, 其数据的空间结构是不同的, 这就导致了不同样本集数据各维度间的相关性存在差异. 针对这一特性, 可以在马氏距离的基础上利用样本的标签信息进行学习以得到距离测度矩阵M, 使得在新的特征空间里同类样本之间的距离尽可能小而异类样本之间的距离尽可能大. 借此, Bar-Hillelet等提出了RCA方法, Hoi等基于改进RCA提出了DCA方法; 在研究已有算法的基础上, 本文提出了一种新的距离测度学习方法I-DCA.

1.2.1　相关分量分析(RCA)

RCA是一种全局性的线性变换方法, 该方法通过将相关性较强的维数赋予较大的权重, 而相关性较弱的维数赋予较小的权重, 来降低数据集的混杂, 以至于在新的特征空间内, 数据的结构更容易被拆散. 为简明算法介绍, 定义“类团”为属于同类别的一系列样本点的集合. RCA算法步骤如下.

(1) 将数据集所包含的样本减去所有样本的均值.

(2) 假设共有 p 个样本形成了 k 个“类团”, 每个“类团”包含n_j个样本, 且均值是 m_j, x_{ji} 是第 j 个“类团”的第 i 个数据, 计算各“类团”经中心化后整个数据集的协方差矩阵C, 公式为

$$C = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-m_j)(x_{ji}-m_j)^{\mathrm{T}}. \tag{2}$$

(3) 计算白化矩阵 $W = C^{-1/2}$, 将其作用于原始数据即 $Y = WX$; 而矩阵 $M = C^{-1}$ 则作为距离测度矩阵用来计算样本间的马氏距离.

1.2.2　差别分量分析(DCA)

针对RCA仅利用同类样本间的对等约束条件而未充分利用异类样本间不对等约束条件这一缺陷, DCA能够最大化不同类数据集差别, 同时最小化同类数据集差别. DCA算法步骤如下.

(1) 计算协方差矩阵 $\hat{C}_b$ 和 $\hat{C}_w$. 定义各参数, D_j 表示第 j 个“类团”的数据集合, $n_b = \sum_{j=1}^{n}|D_j|$ 表示各“类团”中样本数目, $m_j = \frac{1}{n_j}\sum_{i=1}^{n_j}x_{ji}$表示第 j 个“类团”的均值, x_{ji} 是第 j 个“类团”的第 i 个数据. $\hat{C}_b$ 得到不同类数据之间的所有方差, $\hat{C}_w$ 得到同类数据之间的所有方差, 其公式分别为

$$\hat{C}_b = \frac{1}{n_b}\sum_{j=1}^{n}\sum_{i\in D_j}(m_j-m_i)(m_j-m_i)^{\mathrm{T}}, \tag{3}$$

$$\hat{\underset{w}{C}} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}-m_j)(x_{ji}-m_j)^{\mathrm{T}}. \tag{4}$$

(2) DCA的工作在于解决最优化问题, 得到最优化变换矩阵 A, 进而得到最优的距离测度矩阵 $M = A^{\mathrm{T}}A$,

$$J(A) = \arg\max_{A}\frac{|A^{\mathrm{T}}\hat{\underset{b}{C}}A|}{|A^{\mathrm{T}}\hat{\underset{w}{C}}A|}. \tag{5}$$

图2显示了将RCA及DCA算法作用于样例数据集前后的分布情况. 经比较发现, 在新的特征空间内, 同类样本点相对于原始数据更加集中. 由于DCA是RCA基础上的改进, 同时运用了样本数据之间的对等约束和不对等约束, 因此相对于RCA算法, DCA算法对于数据的聚集效果更加明显.

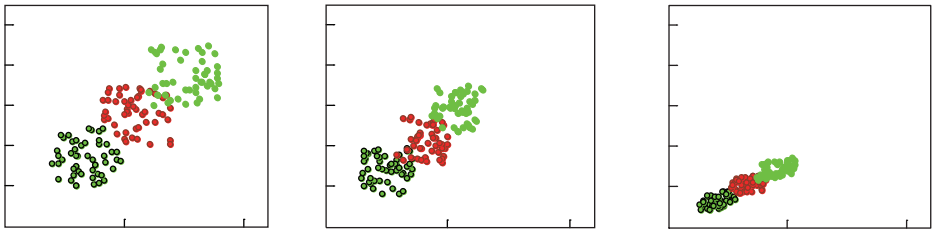

图 2 RCA与DCA算法说明

(a)带有标签信息的原始数据集合; (b)经RCA变换后的数据集合; (c)经DCA变换后的数据集合

Fig. 2 An illustrative example of the RCA and DCA algorithm

(a) The fully labeled data set with 3 classes; (b) The original data after applying the RCA transformation; (c) The original data after applying the DCA transformation.

1.2.3 独立-差别分量分析(I-DCA)

I-DCA方法基于不同类别的样本集其各维度之间的相关性是不一样的这一观点. 首先对各个"类团"分别进行白化变换, 对不同的维数赋予不同的权重使之相互独立, 此时, 同类样本点间差异减小, 不同样本集之间的差异增加. 其次对于包含不同类别样本集的整个数据集来说, 其空间数据结构也发生了变化, 由于属于同类样本集的数据各维度间相互独立, 此时整体数据集各维度的相关性只体现不同样本集之间的差异. 算法步骤如下.

(1) 假设共有 p 个样本形成了 k 个"类团", D_j 表示第 j 个"类团"的数据集合, x_{ji} 是第 j 个"类团"的第 i 个数据, $n_j = \sum_{j=1}^{n}|D_j|$ 表示各"类团"中样本数目, $m_j = \frac{1}{n_j}\sum_{i=1}^{n_j}x_{ji}$ 表示第 j 个"类团"的均值. 计算每一"类团"所包含数据的协方差矩阵 C_j, 公式为

$$C_j = \frac{1}{n_j}\sum_{j=1}^{k}\sum_{i\in D_j}(x_{ji}-m_j)(x_{ji}-m_j)^{\mathrm{T}}. \tag{6}$$

(2) 计算白化矩阵 $W_j = C_j^{-1/2}$, 继而计算 $Y_j = W_jX_j$, 使各"类团"数据各维度相互独立.

(3) 计算由 Y_j 组成的总体样本的协方差矩阵 S, 则该算法的距离测度矩阵为 $M = S^{-1}$.

图3显示了将马氏距离和I-DCA算法运用于样例数据集前后的分布情况. 对比图3(b)和图3(c), 发现不同类别的数据经各自白化矩阵变换后, 样本的分布发生改变, 不同类别的数据

能够相互分离; 对比图3(e)和图3(f)发现, 改进后的算法能够使同类样本数据分布更加紧密, 证明了算法的有效性.

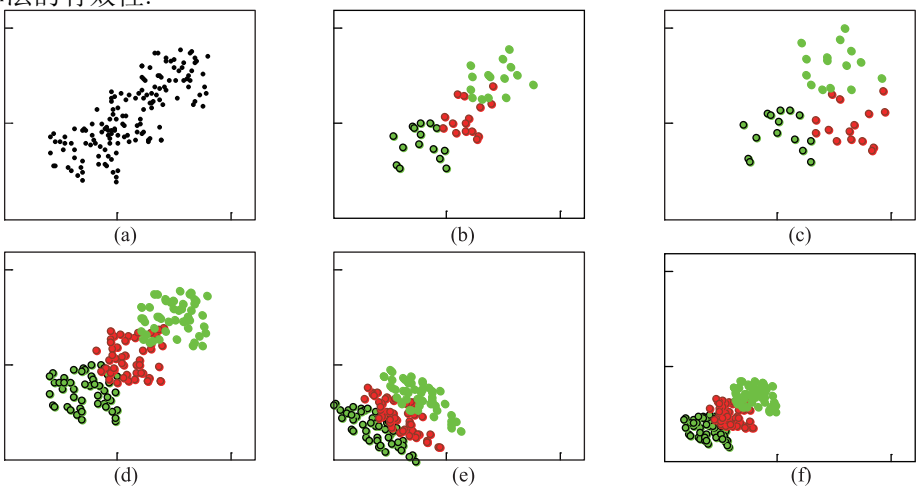

图3　I-DCA算法说明

(a)原始数据集合; (b)随机选择的带有标签信息的训练样本; (c)变换后各维度相互独立的训练样本 (d)带有标签信息的原始数据集; (e)经马氏距离变换后的数据集; (f)经I-DCA算法变换后的数据集

Fig. 3　An illustrative example of the I-DCA algorithm

(a) The fully unlabeled data set; (b) Random labeled train data set; (c) The whitening transformation applied to train data; (d) The fully labeled data set with 3 classes; (e) The original data after applying the mahalanobis distance; (f) The original data after applying the I-DCA transformation.

2　实验结果

2.1 数据描述

实验采用显微高光谱神经采集系统, 对兔子的脊髓前根及脊髓后根的单独未染色切片样本进行采集, 以获取运动神经及感觉神经的透射光谱数据. 鉴于系统中AOTF分光计的性能, 从技术可行性角度选择的光谱范围为可见光区域, 对应的光谱范围为545~894 nm, 频率范围99~178 MHz, 以1 MHz作为频率间隔共采集80个波段的数据[8]. 由于系统中分光计的频率与波长呈非线性关系, 在不同波长处的光谱分辨率也不尽相同, 光谱分辨率在545 nm处为2 nm, 在800 nm处为5 nm. 系统有效像元设定为1 024×1 024, 系统采集的整个视场大小为63 um×63 um, 则其空间分辨率约为0.061 5 um.

对于运动神经及感觉神经的分类, 可以选择神经的轴突以及髓鞘作为两类特征进行分类[12]. 将采集到的显微高光谱图像数据经预处理去除噪声之后, 利用ENVI软件收集运动及感觉神经的轴突以及髓鞘的纯净像元数据作为待分类数据, 数据的类别、名称及数目如表1所示.

表 1　数据类别名称以及数量

Tab. 1　The data type name and quantity

标号	类别名称
1	Motor_axone(2000)
2	Motor_medullary_sheath(2000)
3	Sensory_axone(2000)
4	Sensory_medullary_sheath(2000)

2.2 分类结果

本文实验基于k近邻分类器采用不同距离测度对神经进行分类, 通过统计不同训练数据集下测试数据的相应分类精度以比较不同距离测度的优劣, 共比较了欧氏距离、马氏距离、RCA、DCA以及I-DCA算法下得到的距离测度对于分类的影响, 5种分类结果如表2所示. 为消除实验结果随机性的影响, 重复实验10次取平均值, 每次随机挑选相同比例数目的像元作为训练数据, 余下的像元作为测试数据.

表 2 k近邻总体分类精度比较

Tab. 2 Overall kNN classification precision comparing

	train% | test%	Euclidean distance%	Mahalanobis distance%	I-DCA%	RCA%	DCA%
	10|90	86.5	53.61	86.28	69.92	71.83
k=1	30|70	94.16	59.41	93.87	74.42	77.33
	50|50	96.3	62.42	95.98	75.81	78.31
	10|90	81.72	53.26	81.47	72.07	73.07
k=5	30|70	90.34	61.70	90.26	77.16	78.27
	50|50	93.65	65.09	93.38	79.59	79.53
	10|90	77.66	51.67	77.16	70.81	73.12
k=10	30|70	86.79	58.77	86.63	75.81	77.22
	50|50	90.63	63.81	90.57	77.62	79.32

根据实验结果分析可得以下结果.

(1) 对于神经的纯净像元数据, 除马氏距离外各距离测度下的分类精度都在70%以上, 不仅说明了从光谱角度分类运动神经和感觉神经的可行性, 同时也验证了部分算法的有效性.

(2) 马氏距离是在数据满足高斯分布下假设的, 而神经的显微高光谱数据可能不满足这个条件, 因此其分类效果较差. 而基于马氏距离基础上的RCA、DCA和I-DCA, 因应用了样本的标签信息, 其分类精度确实高于使用马氏距离的分类精度, 且经I-DCA的分类效果最好, 说明了新算法的有效性.

(3) DCA是在RCA的基础上加以改进, 即DCA不仅利用了样本点间的正约束还运用了样本点间的负约束, 使得距离测度矩阵能更好地反应样本间的真实信息. 经实验结果对比, DCA确实比RCA有较好的分类效果.

(4) k值以及训练样本数目的选择也对分类精度有较大影响. 分类精度会随着训练样本数目的增加而有所提高; 但不同的距离测度下, 最高的分类精度却对应着不同的k值, 针对这一点, 相关参数需合理选择.

3 结束语

本文在研究已有的距离测度学习算法的基础上, 针对马氏距离算法未考虑不同类别样本维度间相关性存在差异这一缺陷提出了一种新的算法, 通过将I-DCA、RCA、DCA以及一些传统距离算法应用于运动神经和感觉神经显微高光谱数据分类中, 发现这些算法都具有一定的分类效果, 其中欧氏距离及I-DCA算法的分类精度较高; 相较于其他距离测度学习算法, I-DCA的分类精度较高、算法简单且运算速度较快. 但是为证明算法的广泛性, 后续实验还需将其应用于多种高光谱数据集中.

[参 考 文 献]

[1] 徐沁同. 应用拉曼光谱和超光谱成像技术识别周围神经纤维及神经束功能性质和显微结构的研究[D]. 上海: 复旦大学, 2013.

(下转第 88 页)

[10] PARISI G, PETRONZIO R. On the Breaking of BjorkenScaling [J]. Physics Letters B, 1976, 62(3): 331-334.
[11] GLUCK M, REYA E, VOGT A. Dynamical parton distributions revisited [J]. European Physical Journal C, 1998, 5: 461-470.
[12] CHEN X R, RUAN J H, WANG R, et al. Applications of a nonliner evolution equation Ⅰ: The parton distributions in the proton [J]. International Journal of Modern Physics E, 2014, 23: 1450057.
[13] CHEN X R, RUAN J H, WANG R, et al, Applications of a nonliner evolution equation Ⅱ: The EMC effect [J]. International Journal of Modern Physics E, 2014, 23: 1450058.
[14] CHEN X R, RUAN J H, WANG R, et al. Nucleon spin structure [J]. International Journal of Modern Physics E, 2015, 24: 1550077.
[15] LOU L Y, RUAN J H. A new research about pion parton distribution function [J]. Chinese Physics Letters, 2015, 32(5): 051201.
[16] GRIBOV L V, LEVIN E M, RYSKIN M G. Semihard processes in QCD [J]. Physics Reports, 1983, 100: 1-150.
[17] MUELLER A H, QIU J W. Gluon recombination and shadowing at small values of x [J]. Nuclear Physics B, 1986, 268(2): 427-452.
[18] ZHU W, RUAN J H. A new modified altarelli-parisi evolution equation with parton recombinationin proton [J]. Nuclear Physics B, 1999, 559: 378-392.
[19] ZHU W. A New approach to parton recombination in a QCD evolution equation [J]. Nuclear Physics B, 1999, 551: 245-274.
[20] ZHU W, SHEN Z Q, RUAN J H. Parton recombination effect in polarized partondistributions [J]. Nuclear Physics B, 2004, 692: 417-433.
[21] ZHU W, SHEN Z Q. Properties of the gluon recombination functions [J]. Physics and Nuclear Physics, 2005(2): 109-114.
[22] BADIER J, BOUCROT J, BOUROTTE J, et al. Measurement of the K^-/π^- structure function ratio using the Drell-Yan process [J]. Physics Letter B, 1980, 93: 354-362.
[23] SHIGETANI T, SUZUKI K, TOKI H. Pion structure function in the Nam and Jona-Lasinio model [J]. Physics Letters B, 1993, 308: 383-388.
[24] HOLT R J, ROBERTS C D. Distribution functions of the nucleon and pion in the valence region [J]. Review of Modern Physics, 2010, 82: 2991-3044.
[25] NAM S. Parton-distribution functions for the pion and kaon in the gauge-invariant nonlocal chiral-quark model [J]. Physical Review D, 2012, 86: 074005. DOI: 10.1103/PhysRevD.86.074005.

(责任编辑: 李　艺)

(上接第 74 页)

[2] 房娟, 刘洪英, 陈增淦, 等. 基于显微高光谱成像技术的运动和感觉神经分类研究 [J]. 影像科学与光化学, 2015, 33(3): 203-210.
[3] 刘洪英, 李庆利, 顾彬, 等. 新型分子高光谱成像系统性能分析及数据预处理 [J]. 光谱学与光谱分析, 2012, 32(11): 3161-3166.
[4] 刘博. 距离测度学习理论与应用研究 [D]. 合肥: 中国科学技术大学, 2009.
[5] BAR-HILLEL A, HERTZ T, SHENTAl N, et al. Learning distance function using equivalence relations [C]// Machine Learning, Proceedings of the Twentieth International Conference. 2003: 11-18.
[6] GOLDBERGER J, ROWEIS S T, HINTON G E, et al. Neighbourhood components analysis. [J]. Advances in Neural Information Processing Systems, 2004, 83(6): 513-520.
[7] XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information [J]. Advances in Neural Information Processing Systems, 2003, 15: 505-512.
[8] HOI S C H, LIU W, LYU M R, et al. Learning distance metrics with contextual constraints for image retrieval [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2006: 2072-2078.
[9] 苟建平. 模式分类的 k 近邻方法 [D]. 成都: 电子科技大学, 2012.
[10] 张巍. 基于 k 近邻分类准则的特征变换算法研究 [D]. 上海: 复旦大学, 2007.
[11] 张杰. 基于距离测度学习的图像分类方法研究 [D]. 上海: 复旦大学, 2010.
[12] 刘洪英. 分子超光谱成像的生物组织定量检测与方法研究 [D]. 上海: 华东师范大学, 2011.

(责任编辑: 李　艺)