

基于主动学习的平衡类鉴别分析

黄明晓,荆晓远,李敏,姚永芳

(南京邮电大学 自动化学院,江苏 南京 210003)

摘要:特定类的思想是将传统的多类特征提取和识别任务转化为多个两类问题,由此产生了类不平衡问题,影响最优鉴别特征的提取。为了解决该问题,文中提出了一种主动学习平衡类鉴别分析(ALCBD)方法。对于每个特定类,ALCBD从其对应的大类中选取它的部分近邻样本构成特定类的近邻样本集,接着将这个近邻样本集划分成与特定类相同样本数的多个子集,然后根据主动学习的思想挑选最优子集与特定类结合成为新样本集,最后用传统的线性鉴别分析(LDA)方法得到鉴别向量。基于USPS和Honda/UCSD数据库的实验表明ALCBD方法能够有效地解决类不平衡问题,并改善了识别性能。

关键词:类不平衡;鉴别特征;主动学习;鉴别分析

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)06-0095-04

doi:10.3969/j.issn.1673-629X.2014.06.024

Class-balanced Discriminant Analysis Based on Active Learning

HUANG Ming-xiao, JING Xiao-yuan, LI Min, YAO Yong-fang

(College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: The class-specific idea tends to recast a traditional multi-class feature extraction and recognition task into several binary class problems. In this way, the class-imbalance problem occurs, which might affect the extraction of optimal discriminant features. In order to address this problem, propose an approach named Active-Learning based Class-Balanced Discriminant analysis (ALCBD). For a specific class, ALCBD selects a reduced counterpart class whose data are nearest to the data of specific class, and further divides them into smaller subsets, each of which has the same size as the specific class. Then, ALCBD chooses the optimal subset according to the idea of active learning, and further combines it with the specific class to form a new sample set. Finally perform the Linear Discriminant Analysis (LDA) on them to obtain discriminative vectors. The experimental results on the USPS and Honda/UCSD databases demonstrate that the ALCBD approach can effectively solve the class-imbalance problem, and improve the recognition performance.

Key words: class-imbalance; discriminant features; active learning; discriminant analysis

0 引言

特征提取^[1-2]是模式识别研究中的基本问题之一。目前多数特征提取方法的设计准则是优化整体的识别率,但这样却牺牲了类别的特殊情况,可能会对识别效果造成负面影响。为了解决该问题,Baggenstoss提出了一个想法^[3]:每个类都有自己的特征集,设计概率分类器。基于这个想法,陈晓红等人提出了类特定线性鉴别分析(Class-Specific Linear Discriminant Analysis, CSLDA)^[4]方法。然而,这样的特定类方法由于样本数量的差异,会产生类不平衡问题^[5],难以获得较好的鉴别效果。类不平衡问题在知识挖掘和数据工

程研究中都是个很有挑战的领域^[6]。

主动学习已被很多学者证实是一种很有效的机器学习算法,但将其用于类不平衡的问题却并不多见。主动学习方法可以从样本集中选取部分样本来构造平衡类^[7-8]。为了解决特定类思想所带来的类不平衡问题,提高识别效果,文中提出了一种新方法来从不平衡数据中获取鉴别特征,称为主动学习平衡类鉴别分析(ALCBD)方法。ALCBD以特定类为正类,其余类为负类,构造特定类在负类中的近邻样本集,并将近邻样本集划分成若干份与特定类大小相同的平衡子集,然后从中挑选最有利于主动学习的平衡子集与特定类结

收稿日期:2013-08-13

修回日期:2013-11-18

网络出版时间:2014-02-24

基金项目:国家自然科学基金资助项目(61073113);江苏省普通高校研究生科研创新计划(CXLX13_465)

作者简介:黄明晓(1988-),女,研究生,研究方向为模式识别;荆晓远,教授,博士生导师,研究方向为模式识别、图像与信号处理、信息安全、机器学习与数据挖掘。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0902.027.html>

合成为新样本集,最后用传统的线性鉴别分析(LDA)^[9]方法得到鉴别向量。基于 USPS 手写体数据库^[10]和 Honda/UCSD 人脸数据库^[11]的实验验证了 ALCBD 方法的有效性。

1 主动学习平衡类鉴别分析(ALCBD)

本节详细介绍 ALCBD 方法:首先,对于每个特定类,从其对应的负类中选取它的部分近邻样本构成特定类的近邻样本集;然后,将这个近邻样本集进行随机平衡划分;接着,将划分好的平衡子集与特定类组合成为新样本集,根据主动学习的思想获取最优平衡子集;最后,采用 LDA 方法得到鉴别向量。

1.1 构造特定类的近邻样本集

令 $X_i = \{x_j^i\}_{j=1,2,\dots,n}$ 表示样本数为 n 的第 i 类, $Y_i = \{y_k^i\}_{k=1,2,\dots,(c-1)n}$ 则表示对应的剩余类的样本,剩余类的总大小为 $(c-1)n$, n, c 分别表示一类的样本数和总类别数。将 X_i 看作特定类(即正类), Y_i 为相应的负类。 $D(y_k^i)$ 表示样本 y_k^i 和 X_i 中样本的距离最小值:

$$D(y_k^i) = D_k^i = \min \{ \|y_k^i - x_j^i\| \}_{j=1,\dots,n} \quad (1)$$

从负类中选取最短距离 $\{D_k^i\}_{k=1,2,\dots,(c-1)n}$ 对应的 $\{y_k^i\}_{k=1,2,\dots,(c-1)n}$ 来形成特定类的近邻样本集 Y_i^R 。一般选择 l 个样本来形成特定类的近邻样本集,其中 l 是 n 的倍数。

1.2 特定类的近邻样本集的平衡划分

在特定类近邻样本集确定之后,采用随机划分的方法将特定类的近邻样本集 Y_i^R 划分成 $b(b = l/n)$ 个大小与特定类相等的子集,记第 k 个子集为 $Y_{ik}^R(k = 1, 2, \dots, b)$ 。接下来通过计算总体散度矩阵来估计每个随机划分得到的子集的样本分布情况。

令 \tilde{S}_i^k 表示 Y_{ik}^R 的总体散度矩阵, $\tilde{S}_i^k(k = 1, 2, \dots, b)$ 表示随机划分的第 k 个子集的总体散度矩阵。 \tilde{S}_i^k 的迹, $\tilde{S}_i^k(k = 1, 2, \dots, b)$ 的迹和 $\tilde{S}_i^1, \tilde{S}_i^2, \dots, \tilde{S}_i^b$ 的均值可以用如下的式子表示:

$$\begin{aligned} \text{trace}(\tilde{S}_i^k) &= \text{trace}\left(\frac{1}{M} \sum_{j=1}^M (x_j - \bar{x})(x_j - \bar{x})^T\right) = \\ &= \frac{1}{M} \sum_{j=1}^M [\text{trace}(x_j x_j^T) - \text{trace}(x_j \bar{x}^T) - \\ &\quad \text{trace}(\bar{x} x_j^T) + \text{trace}(\bar{x} \bar{x}^T)] = \\ &= \frac{1}{M} \sum_{j=1}^M x_j^T x_j - \bar{x}^T \bar{x} \quad (2) \\ \text{trace}(\tilde{S}_i^k) &= \text{trace}\left(\frac{1}{n} \sum_{p=1}^n (x_{kp} - \bar{x}_k)(x_{kp} - \bar{x}_k)^T\right) = \\ &= \frac{1}{n} \sum_{p=1}^n [\text{trace}(x_{kp} x_{kp}^T) - \text{trace}(x_{kp} \bar{x}_k^T) - \end{aligned}$$

$$\begin{aligned} &[\text{trace}(\bar{x}_k x_{kp}^T) + \text{trace}(\bar{x}_k \bar{x}_k^T)] = \\ &= \frac{1}{n} \sum_{p=1}^n x_{kp}^T x_{kp} - \bar{x}_k^T \bar{x}_k \quad (3) \end{aligned}$$

$$\begin{aligned} \text{mean}(\text{trace}(\tilde{S}_i^k)) &= \frac{1}{b} \sum_{k=1}^b \left(\frac{1}{n} \sum_{p=1}^n x_{kp}^T x_{kp} - \bar{x}_k^T \bar{x}_k \right) = \\ &= \frac{1}{b} \sum_{k=1}^b \frac{1}{n} \sum_{p=1}^n x_{kp}^T x_{kp} - \frac{1}{b} \sum_{k=1}^b \bar{x}_k^T \bar{x}_k \quad (4) \end{aligned}$$

其中, $M = 2n$ 。

表 1 显示了在 USPS 和 Honda/UCSD 数据库上,随机划分的每个子集 $Y_{ik}^R(k = 1, 2, \dots, b)$ 的总体散度矩阵 $\tilde{S}_i^k(k = 1, 2, \dots, b)$ 的迹,以及 Y_i^R 的总体散度矩阵 \tilde{S}_i^k 的迹,从表中可以看出 $\text{mean}(\text{trace}(\tilde{S}_i^k))$ 的值是与 $\text{trace}(\tilde{S}_i^k)$ 接近的,也就是说,每个随机划分得到的子集 Y_{ik}^R 的样本分布都与整个子集 Y_i^R 的样本分布接近。

表 1 两个数据库上的总体散度矩阵的迹

度量方式	总体散度矩阵的迹	
	USPS($\times 10^2$)	Honda/UCSD($\times 10^9$)
$\text{trace}(\tilde{S}_i^k)$	0.495 4	3.745 9
$\text{mean}(\text{trace}(\tilde{S}_i^k))$	0.498 9 \pm 0.041 1	3.769 6 \pm 0.212 6
$\pm \text{std}(\text{trace}(\tilde{S}_i^k))$		

1.3 基于主动学习的最优平衡子集选择

根据主动学习的思想,学习样本集中最难分类的样本效果最好,即样本越难分类,越利于主动学习。而样本的总体散度越小,聚拢度越高,样本越难分类。因此,下面从随机划分得到的子集中挑选与特定类的样本的总体散度最小的一个子集。

设 d 维训练样本集 $Z_{ik} = \{X_i, Y_{ik}^R\}_{k=1,2,\dots,b}$ 中包含 $M = 2n$ 个训练样本, $Z = \{z_{ik}^1, z_{ik}^2, \dots, z_{ik}^{2n}\}$ 。训练样本的总体散度矩阵可以表示为:

$$S_i^k = \frac{1}{M} \sum_{j=1}^M (z_{ik}^j - m_{ik})(z_{ik}^j - m_{ik})^T \quad (5)$$

这里, m_{ik} 是训练样本集 $Z_{ik} = \{X_i, Y_{ik}^R\}$ 的总均值,即

$$m_{ik} = \frac{1}{M} \sum_{j=1}^M z_{ik}^j \quad (6)$$

由于 $\text{trace}(S_i^k)$ 可以近似地反映训练样本集 Z_{ik} 的总体散度,因此,可以通过对 $\{\text{trace}(S_i^k)\}_{k=1,2,\dots,b}$ 进行降序排序,从中挑选 $\text{trace}(S_i^k)$ 最小的一个子集来构成特定类的最优平衡子集。

1.4 构造投影矩阵

对每一个特定类,在确定了其最优平衡子集 Y_{is}^R 之后,通过结合 X_i 和 Y_{is}^R ,并做 LDA 来为特定类 X_i 学习鉴别特征 W_i 。最后,将所有学习到的向量 W_i 结合到一起,形成整体鉴别变换 $W = [W_1, W_2, \dots, W_c]$ 。

下面给出 ALCBD 算法具体的计算流程:

- 1) 对 $i = 1:C$, 得到每个特定类 X_i 及其对应负类 Y_i 。
- 2) 根据 1.1 构造特定类样本的近邻样本集 Y_i^R 。
- 3) 对特定类的近邻样本集 Y_i^R 进行随机划分, 每个子集的大小与特定类 X_i 相同。
- 4) 根据 1.3 来确定最优的平衡子集 Y_{is}^R , 并构造训练样本 $Z_{is} = \{Y_{is}^R, X_i\}$ 。
- 5) 对 Z_{is} 做 LDA 来获取鉴别向量 W_i 。
- 6) 得到总体鉴别向量 $W = [W_1, W_2, \dots, W_c]$, 并得到变换后的特征 $X' = W^T \cdot X$, 用 cosine 分类器对变换后的特征来分类。

2 实验

本节介绍 ALCBD 方法在 USPS 手写体数据库和 Honda/UCSD 人脸数据库的实验结果, 并将 ALCBD 与传统的 LDA、考虑了样本近邻信息而没有考虑特定类信息的 LFDA^[12-13] 以及考虑了特定类信息却没有考虑类不平衡问题的 CSLDA 进行对比分析。所有的方法均采用基于余弦距离的最近邻分类器来做分类识别。

2.1 数据库介绍

USPS 数据库包含 9 298 个样本, 每个样本是 16×16 像素, 图像是从手写的邮政编码上的数字 0 到 9 收集到的, 这个数据集广泛引用在很多识别方法中。为了简化计算, 文中在每类中选取 200 个样本总计 2 000 个样本形成了实验中的样本集。

Honda/UCSD 人脸视频数据库的每一个视频都是使用 SONY EVID30 照相机在室内环境拍摄的, 每秒 15 帧, 最少 20 秒。每一帧的分辨率都是 640×480 , 每个人都至少有两个视频。文中的实验数据库中包含 20 个人, 每个人从两段视频中选择包含了头部转动的 100 帧, 并截取每一帧中大小为 400×400 的脸部图像。为了减小计算代价, 文中实验将每幅人脸图像压缩到 50×50 。

图 1 和图 2 分别显示了 USPS、Honda/UCSD 数据库中的样本图像。



图 1 USPS 数据库的样本图像

2.2 实验结果及分析

在 USPS 手写体数据库上, 每类随机选取 60 个样本组成训练集, 其余的样本组成测试集; 对每个特定类选择 300 个近邻样本。在人脸数据库上, 每类随机选

取 30 个样本组成训练集, 其余的样本组成测试集; 对每个特定类选择 150 个近邻样本。实验都采取随机挑选训练样本的方式运行 20 次。



图 2 Honda/UCSD 数据库的样本图像

图 3 和图 4 分别给出了所有方法在 USPS 和 Honda/UCSD 两个数据库上随机 20 次的识别率波动图。表 2 给出了相应的平均识别率。从表 2 可以看出, 在 USPS 手写体数据库上, 与传统的没有考虑到特定类信息的 LDA 相比, ALCBD 方法具有明显优势, 相较于考虑了样本近邻信息的 LFDA, ALCBD 的平均识别率高 4.17% (= 88.62% - 84.45%), ALCBD 的平均识别率比未考虑类不平衡影响的 CSLDA 高 3.68% (= 88.62% - 84.94%)。在 Honda/UCSD 人脸数据库上, ALCBD 的平均识别率比 LDA 高 4.45% (= 96.62% - 92.17%), 比 LFDA 高 2.39% (= 96.62% - 94.23%), 比 CSLDA 高 1.5% (= 96.62% - 95.12%)。

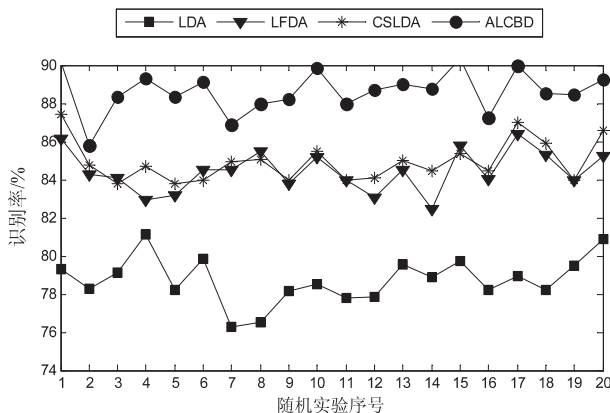


图 3 所有方法在 USPS 数据库随机 20 次的识别率

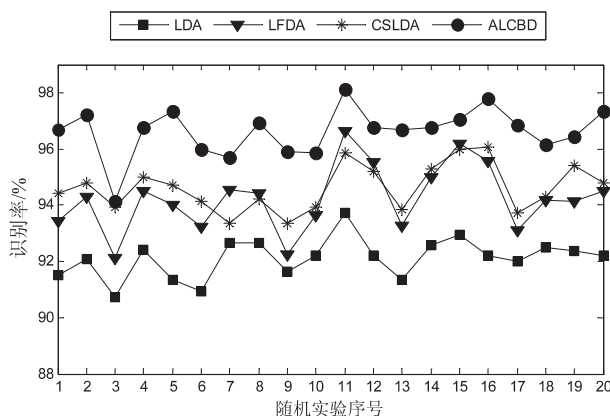


图 4 所有方法在 Honda/UCSD 数据库随机 20 次的识别率

表 2 所有方法的平均识别率

方法名称	平均识别率/%	
	USPS 库	Honda/UCSD 库
LDA	77.23	92.17
LFDA	84.45	94.23
CSLDA	84.94	95.12
ALCBD	88.62	96.62

3 结束语

为了解决特定类思想所带来的类不平衡问题,文中提出了一种新方法从不平衡数据中获取更有效的鉴别特征,称为主动学习平衡类鉴别分析(ALCBD)方法。基于 USPS 和 Honda/UCSD 数据库上的实验表明,文中提出的 ALCBD 方法与 LDA、LFDA 和 CSLDA 三种相关方法相比,有效地提高了识别性能。

参考文献:

[1] 赵振勇,王保华,王 力,等. 人脸图像的特征提取[J]. 计算机技术与发展,2007,17(5):221-224.

[2] 韩 璐. 一种基于 2DLPP 和 2DLDA 的人脸识别方法研究[J]. 计算机技术与发展,2012,22(9):87-90.

[3] Baggenstoss P. Class-specific feature sets in classification[J]. IEEE Trans on Signal Processing,1999,47(12):3428-3432.

[4] 陈晓红,陈松灿. 类依赖的线性判别分析[J]. 小型微型计算机系统,2008,29(5):894-897.

[5] Japkowicz N,Stephen S. The class imbalance problem;a systematic study[J]. Intelligent Data Analysis,2002,6(5):429-

449.

[6] He Haibo,Garcia E A. Learning from imbalanced data[J]. IEEE Trans on Knowledge and Data Engineering,2009,21(9):1263-1284.

[7] Ertekin S,Huang Jian,Bottou L,et al. Learning on the border: active learning in imbalanced data classification[C]//Proc of the sixteenth ACM conf on information and knowledge management. New York,NY,USA:ACM,2007:127-136.

[8] Wang Jinghua,You J,Li Qin,et al. Extract minimum positive and maximum negative features for imbalanced binary classification[J]. Pattern Recognition,2012,45(3):1136-1145.

[9] Belhumeur P N,Hespanha J P,Kriegman D J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,1997,19(7):711-720.

[10] Ma Zhanyu,Leijion A. Bata mixture models and the application to image classification[C]//Proceedings of the 16th IEEE international conf on image processing. Cairo:IEEE,2009:2045-2048.

[11] Lee K C,Ho J,Yang M H,et al. Visual tracking and recognition using probabilistic appearance manifolds[J]. Computer Vision and Image Understanding,2005,99(3):303-331.

[12] Sugiyama M. Local Fisher discriminant analysis for supervised dimensionality reduction[C]//Proc of the 23rd international conf on machine learning. New York,NY,USA:ACM,2006:905-912.

[13] Sugiyama M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis[J]. Journal of Machine Learning Research,2007,8:1027-1061.

(上接第 94 页)

京:南京邮电大学,2013.

[4] 王 钰. RFID 防碰撞算法研究[D]. 南京:南京邮电大学,2011.

[5] Yu Songsen,Zhan Yiju,Wang Yonghua. RFID anti-collision algorithm based on bi-directional binary exponential index[C]//Proceedings of the IEEE international conference on automation and logistics. Jinan:IEEE,2007:2917-2921.

[6] Zhao Jie,Wu Lining. The improvement of RFID anti-collision algorithm[C]//Proceedings of the 24th Chinese control and decision conference. Taiyuan:[s. n.],2012:3261-3264.

[7] 单承赣,余春梅,王聪聪. 改进的二进制查询树的 RFID 标签防碰撞算法[J]. 合肥工业大学学报(自然科学版),2008,31(11):1801-1804.

[8] Kim S,Kim Y,Choi W,et al. A tag prediction anti-collision algorithm using extra bits for RFID tag identification[J]. International Journal of Ad Hoc and Ubiquitous Computing,2012,10(3):164-174.

[9] Galiotto C,Marchetti N,Prasad N,et al. Low access delay anti

-collision algorithm for readers in passive RFID systems[J]. Wireless Personal Communications,2012,64(1):169-183.

[10] 王 荃,滑 楠,张 璐,等. 基于四元查询树算法的改进防碰撞算法[J]. 空军工程大学学报(自然科学版),2012,13(6):75-79.

[11] Wang Tsan-pin. Enhanced binary search with cut-through operation for anti-collision in RFID systems[J]. IEEE Communications Letters,2006,10(4):236-238.

[12] Konstantinou N. Expowave: an RFID anti-collision algorithm for dense and lively environments[J]. IEEE Transactions on Communications,2012,60(2):352-356.

[13] Djeddou M,Khelladi R,Benssalah M. Improved RFID anti-collision algorithm[J]. International Journal of Electronics and Communications,2013,67(3):256-262.

[14] 张志涌,杨祖樱. MATLAB 教程 R2011a[M]. 北京:北京航空航天大学出版社,2011.

[15] 周晓光,王晓华,王 伟. 射频识别(RFID)系统设计与仿真与应用[M]. 北京:人民邮电出版社,2008.

基于主动学习的平衡类鉴别分析

作者:

[黄明晓](#), [荆晓远](#), [李敏](#), [姚永芳](#), [HUANG Ming-xiao](#), [JING Xiao-yuan](#), [LI Min](#),
[YAO Yong-fang](#)

作者单位:

[南京邮电大学 自动化学院, 江苏 南京, 210003](#)

刊名:

[计算机技术与发展](#) 

英文刊名:

[Computer Technology and Development](#)

年, 卷(期):

[2014\(6\)](#)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201406024.aspx