

引用格式:叶丹,李智,王勇军. 基于 SPLDA 降维和 XGBoost 分类器的行为识别方法研究[J]. 微电子学与计算机,2019,36(6):35-39. [YE D,LI Z,WANG Y J. Research on behavior identification based on SPLDA dimensional reduction algorithm and XGBoost classifier [J]. Microelectroncs & Computer,2019,36(6):35-39.]

基于 SPLDA 降维和 XGBoost 分类器的行为识别方法研究

叶 丹¹,李 智²,王勇军^{1,2}

(1 桂林电子科技大学 电子工程与自动化学院,广西 桂林 541004;

2 桂林航天工业学院 无人机遥测重点实验室,广西 桂林 541004)

摘 要: 针对人体行为识别过程中分类算法识别精度低和数据样本集的“维数灾难”问题,提出了基于行为识别的 SPLDA 降维算法. 首先,利用 SPLDA 算法在原有样本协方差矩阵不变的情况下获取最重要的主分量,通过贪婪搜索方法得到多个投影向量;然后,通过更新类内散度矩阵获得最优转换矩阵;最后,将降维后的样本数据集通过 XGBoost 分类器进行最终的行为识别. 实验结果表明,XGBoost 分类器与随机森林算法相比,平均识别精度提高了 2.66%,识别时间降低了 0.52 s;SPLDA-XGB 算法可以实现有效降维且比 PCA 算法、LDA 算法、LPP 算法、L-PCA 算法与 XGBoost 分类器结合的识别算法具有更高的人体行为识别准确率.

关键词: 行为识别;SPLDA;投影向量;降维算法;分类

中图分类号: TP301.6

文献标识码: A

文章编号: 1000—7180(2019)06—0035—05

Research on behavior identification based on SPLDA dimensional reduction algorithm and XGBoost classifier

YE Dan¹,LI Zhi²,WANG Yong-jun^{1,2}

(1 School of Electronic Engineering and Automation,Guilin University of Electronic Technology, Guilin 541004, China;2 Key Laboratory of Unmanned Aerial Vehicle Telemetry,Guilin University of Aerospace Technology,Guilin 541004, China)

Abstract: Aiming at the problem of "dimension disaster" in human behavior recognition and the classification algorithm has low recognition accuracy and data sample set. First, the SPLDA algorithm is used to obtain the most important principal components with the original sample covariance matrix unchanged, and multiple projection vectors are obtained by greedy search method. Then, the optimal transformation matrix is obtained by updating the class inner divergence matrix. Finally, the dimensionally reduced sample data set is identified by the XGBoost classifier. Experimental results show that compared with the random forest algorithm, the average recognition accuracy of XGBoost classifier is improved by 2.66% and the recognition time is reduced by 0.52s. SPLDA-XGB algorithm can achieve effective dimensionality reduction and has higher accuracy rate of human behavior recognition than PCA algorithm, LDA algorithm, LPP algorithm, l-pca algorithm combined with XGBoost classifier.

Key words: behavior recognition; SPLDA; projected vector; dimensionality reduction algorithm; classification

收稿日期: 2018—08—14; 修回日期: 2018—09—05

基金项目: 广西自然科学基金重点项目(2016GXNSFDA380031);广西自动检测技术与仪器重点实验室基金项目(YQ14203);广西高校无人机遥测重点实验室开放基金项目(WRJ2016KF01)

1 引言

人体行为识别是计算机视觉与模式识别领域的热门研究方向,其通过对人体的动作类型和行为模式进行分析和识别,并用自然语言表达出来^[1]. 目前主流的人体行为识别方式有两种:第一种为基于计算机视觉的行为辨识^[2];第二种为基于传感器技术的行为识别. 利用传感器技术进行行为识别的方式,因其经济性和便携性受到了越来越多的关注^[3]. 基于此,本文选择基于智能手机的加速度传感器和陀螺仪作为采集人体行为数据的信息来源.

在模式识别过程中,高维数据造成存储成本上升的同时降低了分类算法的识别能力,容易出现“维数灾难”问题^[4]. 高维样本数据集只有从维数上进行合理约减,去掉数据集中大部分不相关属性,才能从根本上降低干扰数据对识别算法的影响,提高算法精度.

本文提出的 SPLDA 算法能够保证在原有样本协方差结构不变的情况下获得变换矩阵中最重要的主成份,通过调节类间散度矩阵和类内散度矩阵,使得类间距离最大化的同时类内距离最小化^[5],通过贪婪搜索法获得多个最优投影向量,更新类内散度矩阵得到最佳转换矩阵,提高算法识别率.

2 SPLDA 算法

数据分析方法中的经典 PCA 算法是一种无监督降维算法,将多维原始样本数据集通过主成份分析投影到低维线性子空间来实现数据降维. 在 PCA 算法选择最佳投影矩阵的过程中,没有考虑数据类别的相关信息,在实际应用中,会因为数据投影结果的重叠造成线性不可分的情况. LDA 算法采用的是有监督降维,将样本数据集的每个样本有类别输出,投影后使得类内方差最小、类间方差最大,进而解决其线性不可分的情况. 本文算法是在 PCA 模型的基础上,引入改进的 LDA 算法,提出 SPLDA 算法来实现样本数据集的降维.

原始样本数据集为 X ,其在低维空间的数据表示为 Y ,降维目标函数可以表示为:

$$\begin{aligned} U_{\text{SPLDA}} &= \arg \max \sum_{i=1}^n \vec{y}_i - \vec{y}^m \\ &= \arg \max \sum_{i=1}^n U_{\text{SPLDA}}^T (\vec{x}_i - \vec{x}^m)^2 \\ &= \arg \max \text{tr} (U_{\text{SPLDA}}^T S_T U_{\text{SPLDA}}) \end{aligned} \quad (1)$$

式中, $\vec{y}^m = \frac{1}{n} \sum_{i=1}^n \vec{y}_i$, $\vec{x}^m = \frac{1}{n} \sum_{i=1}^n \vec{x}_i$, S_T 为全局

离散矩阵, S_T 表示如下:

$$S_T = \sum_{i=1}^n (x_i - \bar{x}) (x_i - \bar{x})^T \quad (2)$$

当转换矩阵在尺度上满足 S_T (其中 I_d 为 d 行 d 列的单位矩阵),目标函数可以转换为 S_T 标准特征值的求解问题. 因此,求其低维目标样本数据集即求全局离散矩阵 S_T 中 d 个最大特征值对应的 d 个 m 维特征向量^[6]. 为了避免投影之后不同类别的数据混合,引入类间离散矩阵 D_B 和类内离散矩阵 D_W :

$$S_T = W_1 D_B + W_2 D_W \quad (3)$$

式中, W_1 和 W_2 为加权因子, $W_1 + W_2 = 1$. LDA 降维算法的优化形式可以为:

$$\arg \max \frac{\text{tr}(U_{\text{SPLDA}}^T D_B U_{\text{SPLDA}})}{\text{tr}(U_{\text{SPLDA}}^T D_W U_{\text{SPLDA}})} \quad (4)$$

上式降维算法的求解等同于 $D_B W = \lambda D_W W$ 的特征分解. 此时, λ 为常数,若 W 可逆,则 $D_B = \lambda D_W$,则可得到公式

$$S_T = (W_1 \lambda + W_2) D_W \quad (5)$$

由以上公式可知,算法的实现最终可以转化为求类内散度矩阵中 d 个最大特征值对应的 d 个 m 维特征向量.

在 LDA 算法中引入 $L1$ 范数,目标函数可以表示为:

$$U_{\text{SPLDA}}(w) = \frac{\sum_{i=1}^c N_i ||W^T(\vec{x}_i - \vec{x})||}{\sum_{i=1}^c \sum_{j \in C_i} ||W^T(x_j - \vec{x}_i)||} \quad (6)$$

为了减少数据冗余,将投影矩阵 W 限制为单位正交阵,并且根据 $L1$ 范数的定义,可以将公式(6)转化为绝对值形式:

$$U_{\text{SPLDA}}(w) = \frac{\sum_{i=1}^c N_i \sum_{k=1}^n |W_k^T(\vec{x}_i - \vec{x})|}{\sum_{i=1}^c \sum_{j \in C_i} \sum_{k=1}^n W_k^T(x_j - \vec{x}_i)} \quad (7)$$

将上式目标函数的优化问题转换化为 $n = 1$ 时通过贪婪搜索方法求解多个 ($n > 1$) 投影向量的问题.

当 $n = 1$ 时,目标函数为:

$$\begin{aligned} U_{\text{SPLDA}}(w) &= \frac{\sum_{i=1}^c N_i ||W^T(\vec{x}_i - \vec{x})||}{\sum_{i=1}^c \sum_{j \in C_i} ||W_j^T - \vec{x}_i||} \\ &= \frac{\sum_{i=1}^c N_i |W^T(\vec{x}_i - \vec{x})|}{\sum_{i=1}^c \sum_{j \in C_i} |W_j^T - \vec{x}_i|} \end{aligned} \quad (8)$$

式(9)通过贪婪搜索方法可以得到一个较好的近似最优解.对于局部最优解,则通过迭代算法来获取.假设 W_p 为投影向量,即局部最优解,则 $W(t)$ 表示第 t 次迭代的结果.那么可以将迭代算法的迭代过程作如下描述:

(1) 初始化 $W(t)$,归一化处理 $W(0)$.设 $t = 0$,则 $W(0)$ 的计算公式如下

$$W(0) = \arg \max ||\vec{x_i} - \vec{x}||_2 \quad (9)$$

(2) 定义两个极性函数后带入目标函数:

$$U_{\text{SPLDA}}(w(t)) = \frac{\sum_{i=1}^C N_i P_i(t) W^T(t) (\vec{x_i} - \vec{x})}{\sum_{i=1}^C \sum_{j \in C_i} q_j(t) W^T(t) (x_j - \vec{x_i})} \quad (10)$$

(3) 定义 $d(t)$

$$d(t) = \frac{\sum_{i=1}^C N_i P_i(t) (\vec{x_i} - \vec{x})}{\sum_{i=1}^C N_i |W^T(t) (\vec{x_i} - \vec{x})|} - \frac{\sum_{i=1}^C \sum_{j \in C_i} q_j(t) (x_j - \vec{x_i})}{\sum_{i=1}^C \sum_{j \in C_i} |W^T(t) (x_j - \vec{x_i})|} \quad (11)$$

(4) 更新 $w(t+1)$

$$W(t+1) = W(t) + \beta d(t) \quad (12)$$

对 $w(t+1)$ 进行归一化处理,并且令 $t = t+1$.其中 β 为步长参数,给其设置为一个接近于 0 但是大于 0 的数,可以通过多次实验选取使得目标函数最大时的 β 值.从而确定第 t 次的迭代结果, $w(t+1)$ 为沿着目标函数梯度上升的方向更新.

(5) 收敛性检查.若 $U_{\text{SPLDA}}(W(t+1))$ 没有明显增长,表明迭代结束,输出的即为投影向量;否则,迭代继续^[7].

定理:在迭代过程中,目标函数是关于迭代次数的非递减函数^[8].

为了获取数据集更多的投影向量,基于上述迭代法扩展到多个最优投影向量:

① 初始化 $W(0) = 0, (\vec{x_i} - \vec{x})_0 = \vec{x_i} - \vec{x}$,

$(x_j - \vec{x_i})_0 = x_j - \vec{x_i} (i = 1, \dots, \text{Cand} j \in C_i)$

② 学习 n 个局部最优投影向量:

$(\vec{x_i} - \vec{x})_n = (\vec{x_i} - \vec{x})_{n-1} - W_{n-1} W_{n-1}^T (x_j - \vec{x_i})_{n-1}$,

以 $(\vec{x_i} - \vec{x})_n$ 和 $(x_j - \vec{x_i})_n$ 构成训练集,运用迭代法求 W_n .

定理:通过贪婪搜索法学习的 n 个局部最优投

影向量是单位正交的.

$$D_W = \frac{1}{N} \sum_{i=1}^C \sum_{j \in C_i} W^T (x_j - \vec{x_i})^T W \quad (13)$$

根据投影向量更新 D_W ,求 D_W 中 d 个最大特征值对应的 d 个 m 维特征向量即为最佳转换矩阵.

3 XGBoost 分类器

XGBoost 分类器是 Boosting 方法一种极好的实现,是由陈天奇博士在 2014 年首次正式提出^[9],该分类器是 GBDT (Gradient Boost Decision Tree) 分类器基础上演化而来. XGBoost 分类器的分类是由学习模型、参数调节和最优化目标函数组成.其中学习模型是基函数和权重组合的结果,参数调节是建立模型和优化模型的必要过程,目标函数的优化程度决定了最终模型学习的准确率.目标函数优化的越好,模型泛化能力也就越好.因此,通过最小化损失函数和增加模型复杂度的惩罚项来实现对目标函数的优化,实现模型复杂度和误差综合最优.给定目标函数为:

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta) \quad (14)$$

式中, $L(\theta)$ 为误差函数; $\Omega(\theta)$ 为复杂度函数.分类器的目的是在使得损失偏差较小的情况下函数的复杂度尽可能降低.为了避免过拟合和欠拟合的问题,可以通过在方差和偏差之间寻找一个平衡点,使得学习的模型最优.

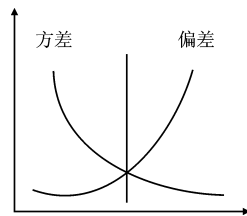


图1 方差偏差图

如图 1,最优点为方差与偏差的相交点.为了使模型的预测能力达到最优,寻找方差与偏差同时处于相对较低点时的平衡点作为最优点.因此,将 XGBoost 分类器目标函数结合方差与偏差可以表示为如下形式:

$$\text{Obj}(\theta) = \sum_i l(y_i + y_i) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T ||W||_j^2 \quad (15)$$

按照梯度下降算法得到弱分类器,然后将每个弱分类器相加,反复计算目标函数的误差,使得最终目标函数的误差最小,则得到的分类器必然最优.

对于 XGBoost 分类器性能的验证,本文选择了

UCI 机器学习库中的 iris 数据集通过与随机森林 (RF) 算法进行对比验证,可以得出:在时间上,XGBoost 分类器比随机森林算法减少了 0.52 s,分类

精度上分别提升了 4.66%,2%,1.33%,2.66%。如图 2 为随机林算法的分类结果,图 3 为 XGBoost 分类器的分类结果。

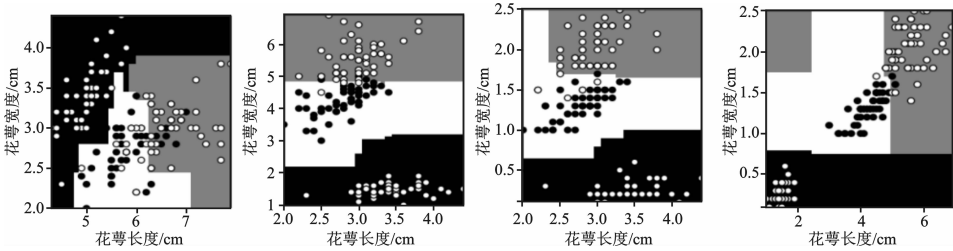


图 2 随机森林算法分类结果

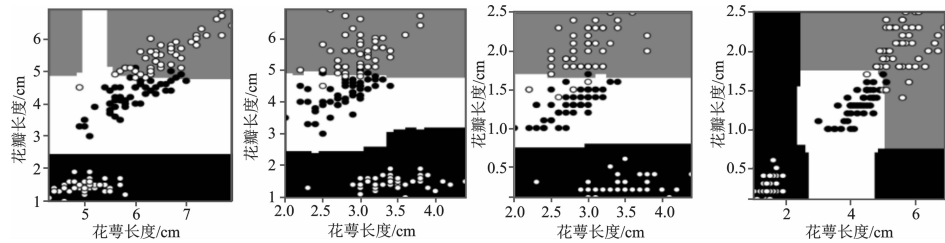


图 3 XGBOOST 算法分类结果

4 实验

4.1 实验数据

本文中样本数据来源于 UCI 机器学习库中 HAR 行为识别数据集,该数据集是通过 Samsung Glalxy SII 智能手机上的加速度传感器和陀螺仪分别采集 30 名志愿者上楼、下楼、坐、站、走、躺六种行为动作的加速度信号和角速度信号,将采集到的信号通过 2.65 s 固定宽度的滑动窗口以及 50%数据重叠的滑动窗口进行信号预处理,计算变量的时频域得到均值、标准差、熵、信号相对性、信号幅值等用于实验的 17 个信号.算法验证是基于 Python2.7 处理平台和 Pycharm 编辑器。

4.2 实验验证及分析

对于人体坐、走、上楼、下楼、站、躺六种行为的分类,通过 SPLDA 算法进行特征约减,去除数据的冗余信息,保留数据集的主要成份信息,然后通过 XGBoost 分类器进行六种行为的分类,行为分类流程图如图 4 所示。

为了验证本文提出的降维算法的准确率,分别使用 SPLDA-XGB 算法与仅使用 XGBoost 分类器对 HAR 数据集进行验证.由实验结果如图 5,在上楼、坐这两个动作 SPLDA-XGB 算法比 XGBoost 分类器识别率分别提高了 4.07%,2.47%;对下楼的行为,本文的算法较 XGBoost 分类器提高了

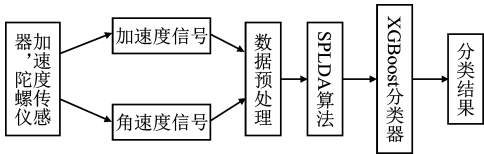


图 4 行为识别流程图

7.69%,由此可以看出,本文所提出的特征约减算法提高了行为识别的准确率。

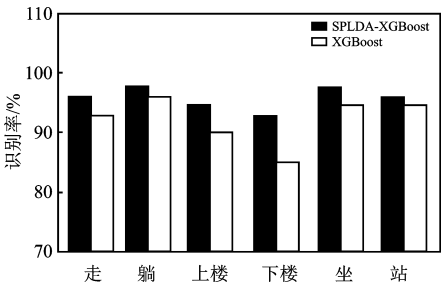


图 5 实验中算法的识别率

运用 SPLDA 算法对 HAR 标准数据集的维度进行特征约减,基于加速度传感器和陀螺仪产生的信号经过预处理之后得到的 7767 * 561 的数据集进行降维,特征数量从 561 维降到了 86 维.为了评估 SPLDA 降维算法的优越性,在选用 XGBoost 分类器的基础上,通过与其他经典的降维算法进行对比.经过 SPLDA 算法降维处理后的数据集比其他四种算法用于 XGBoost 分类器后获得的准确率更高。

由实验结果可知,SPLDA-XGB 算法优于其他算法.特别是躺和坐的行为,本文的算法具有更优的准确率.

表 1 不同特征约减方法的识别率

	PCA-XGB	LDA-XGB	LPP-XGB	L-PCA-XGB	SPLDA-XGB
走	88.54	89.76	95.23	91.17	96.27
躺	91.13	94.18	96.42	95.24	97.86
上楼	83.77	88.34	92.26	89.34	94.17
下楼	81.23	86.16	91.52	88.34	92.16
坐	85.68	92.13	95.27	89.87	97.51
站	94.32	95.16	96.12	93.17	96.58
均值	87.45	90.96	94.47	91.19	95.76

如表 1 所示,对于站这一行为,5 种算法取得了差不多的效果;对于躺的行为,LDA-XGB 算法和 PCA-XGB 算法效果较差;上楼和下楼的行为中,仅有 LPP-XGB 算法和 SPLDA-XGB 算法取得了不错的精度,且 SPLDA-XGB 算法比 LPP-XGB 算法分别提高了 1.91%和 0.64%的准确率;对于坐的行为,SPLDA-XGB 算法准确率达到 了 97.51%,远高于其他算法.

5 结束语

SPLDA 降维算法继承了 PCA 算法和 LDA 算法的优点,通过引入 L1 范数、贪婪搜索法和迭代法来进一步提高样本数据集降维准确率.XGBoost 分类器分类精度高,分类速度快,将 SPLDA 算法与 XGBoost 分类器用于人体行为识别分类,实验证明对于六种人体行为的平均识别率达到 95.76%的识别精度.因此本文提出的 SPLDA-XGB 分类识别算法优于 PCA-XGB、LDA-XGB、LPP-XGB、L-PCA-XGB 算法,能够适用于高维大规模数据集.

参考文献:

[1] KHAN A M,TUFAIL A,KHATTAK A M. Ativity recognition on smartphones via sensor-fusion and KDA-based SVMs[J]. International Journal of Distributed Sensor Networks,2014(1):1-14.

[2] KWAPIAZ J R,WEISS G M,MOORE S A. Activ-ity recognition using cell phoneaccele ro-meters[J]. ACM SigKDD Explorations Newsle-tter,2011,12(2):74-82.

[3] YANG J,WRIGHT J,HUANG T S,et al. Image s-up-erresolution via sparse representation[J]. IEEE Transactions on Image ProcessingA Publication of the IEEE Signal Pocessing So-ciety,2010,19(11):28-128.

[4] 聂军. 基于 K-L 特征压缩的云计算冗余数据降维算法[J]. 微电子学与计算机,2016,33(2):125-129.

[5] VAN D M L,POSTMA E,VAN DENIIERIK J. Di-mension reduction: A comparative review[R]. TiC-CTR,Tiburg University, 2009.

[6] 李龙龙,何东健,王美丽. L-PCA 算法下的高维图像降维算法研究[J]. 西安科技大学学报,2017,37(6):906-911.

[7] 钟福金. 鲁棒线性子空间学习算法与框架研究[D]: 成都:西南交通大学,2015.

[8] Chen T Q. XGBoost: A Scalable Tree Boosting System [C]. In Proceeding KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016:2-4.

[9] Wang H X,Qin T, Zheng W M. L1-norm-based com-mon spatial patterns. [J]. IEEE Transactions on Bio-medical Engineering,2012,59(3):653-662.

作者简介:

叶 丹 女,(1991-),硕士研究生. 研究方向为模式识别与智能系统. E-mail:yedanguet@163. com.

李 智 男,(1965-),教授,博士生导师. 研究方向为智能仪器系统、现代测试理论与技术.

王勇军 男,(1985-),讲师,博士研究生. 研究方向为无人机组合导航技术.

(上接第 34 页)

[8] WORKGROUPO. Open NANDFlash interface specifi-cation[Z]. Revision, 2014, 2: 1-174.

[9] IM J-W, JEONG W, KIM D-H, et al. A 128Gb 3b/cell V-NAND Flash memory with 1Gb/s I/O rate [C]// International Solid-State Circuits Conference, USA, 2015:1-3.

作者简介:

王美兰 女,(1993-),硕士研究生. 研究方向为数字集成电路设计. E-mail:wangmeilan16@mails. ucas. ac. on.

王 颀 男,(1975-),博士,研究员,博士生导师. 研究方向为新型存储器设计技术研究.