第15卷 第29期 2015年10月
1671—1815(2015)29-0062-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 15 No. 29 Oct. 2015

基于 PCA 改进的快速 Adaboost 算法研究

袁 双[1,2] 吕赐兴[1]

(中国科学院沈阳自动化研究所[1],沈阳 110016;中国科学院大学[2],北京 100049)

摘 要 针对传统的 Adaboost 算法可能出现在应对较大训练数据集训练时间过长的问题,提出了一种改进的 Adaboost 算法——PCAdaboost。改进算法利用 PCA 方法的降维技术,对训练样本特征提取主要成分,去除输入样本特征间的相关性,提高分类精度。同时,从样本阈值搜索角度考虑了特征值等分和特征值空间维数,给出了阈值快速搜索方法。实验结果表明,该算法在 UCI 数据集上取得较好的效果。

关键词 PCAdaboost 主成分 阈值搜索 降维

中图法分类号 TP301.6; **文献标志码** A

Boosting,也称为提升或增强学习方法,是一种与 bagging 很类似的集成分类器方法,它的思想起源于 Valiant 提出的 PAC (probably approximately correct) 学习模型[1]。同时,Kearns 等[2]首次提出了 PAC 学习模型中弱学习算法是否能够提升为强学习算法,即两者的等价性问题。1990 年,Schapire[3]最先构造出一种多项式时间内完成的学习算法,对该问题做了肯定的证明,这就是最初的 Boosting 算法。1991 年,Freund[4]对上述算法进行改进,提出了一种更高效率的 Boosting 算法。这两个算法在实践中都存在不可忽视的缺陷,那就是都要求事先知道弱学习算法学习正确的下限。1995 年,Freund 和 Schapire[5]提出了 AdaBoost (adaptive boosting) 算法,该算法拥有扎实的理论基础,并且结构简单,应用广泛,被评为数据挖掘十大算法之一。目前,Adaboost 已经在很多领域得到了广泛应用,如人脸识别[6]、目标跟踪[7]、电力负荷预测[8]。

传统的 Adaboost 算法在面对训练样本和特征较多时,训练时间较长。针对上述问题,贾慧星等[9]提出动态权重裁剪的 Adaboost 训练算法。但该算法存在一定的局限性:裁减系数选择过大,会出现提前退出的情况。对于数据集中存在噪声和特殊样本集的情况,可能无法正确分类导致样本权重过适应。

针对上述问题,首先利用 PCA 方法的降维技术,对训练样本特征提取主要成分,去除输入样本特征间的相关性,并提出了一种基于动态步进搜索的阈值求解方式,大大提高了训练速度。实验结果显示,本文提出的基于 PCA 改进的快速 Adaboost 训练算法,在训练速度和精度上都有很好的效果。

2015 年 6 月 18 日收到 国家高技术研究发展计划(2015AA042102) 资助

第一作者简介:袁 双(1990—),女,汉族,山东淄博人,硕士研究生。研究方向:模式识别。E-mail:yuanshuang@sina.cn。

1 PCA 方法

主成分分析(PCA)是一种非常有用的数学方法,通过正交变换,利用二阶的统计信息进行计算。它是一种最小均方意义上的最优变换[10]。作为一种正交变换,一方面去除输入随机向量之间的相关性,突出原始数据中的隐含特性从而提高分类精度。另一方面是将多属性信息转换为少数几个主成分,这几个主成分包含了大量的属性信息,可以提高分类的效率。K-L 变换理论是 PCA 算法的主要技术手段。

1.1 K-L 基本变换[11]

假设 $\boldsymbol{X}$ 是 n 维随机向量,可以表示成 n 个 n 维向量的加权组合:

$$\boldsymbol{X} = \sum_{i=1}^{n} \alpha_i \varphi_i = \boldsymbol{\Phi a} \tag{1}$$

式(1)中 $\boldsymbol{a} = (\alpha_1, \alpha_2, \cdots, \alpha_n)^{\mathrm{T}}$ 为加权系数,$\boldsymbol{\Phi}$ 为 n 维的向量,即

$$\boldsymbol{\Phi} = (\varphi_1, \varphi_2, \cdots, \varphi_n) \tag{2}$$

由 K-L 变换原理可知,向量 $\boldsymbol{\alpha}$ 中的各个向量应该是互不相关的,如何选取向量 $\boldsymbol{\varphi}$ 保证向量 $\boldsymbol{\alpha}$ 的各个向量互不相关呢?假设向量 $\boldsymbol{\alpha}$ 的各个向量互不相关,即满足式(3)的关系。

$$\alpha_i \alpha_j = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases} \tag{3}$$

即

$$\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}} = \boldsymbol{D}_{\lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{4}$$

假设 $\boldsymbol{\Phi}$ 为归一正交矩阵,即

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} = \boldsymbol{I} \tag{5}$$

将式(1)两边左乘 $\boldsymbol{\Phi}^{\mathrm{T}}$,根据 $\boldsymbol{\Phi}$ 的正交性,可知 K-L 变换式

$$\boldsymbol{a} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x} \tag{6}$$

假设向量 $\boldsymbol{X}$ 的自相关矩阵 $\boldsymbol{R}$ 为式(7),将式(1)代入式(6),得

$$\boldsymbol{R} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} = \boldsymbol{\Phi}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}} = \boldsymbol{\Phi}\boldsymbol{D}_{\lambda}\boldsymbol{\Phi}^{\mathrm{T}} \tag{7}$$

式(5)中 $\boldsymbol{I}$ 为单位矩阵,将式(5)右乘 $\boldsymbol{\Phi}$,得

$$\boldsymbol{R}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{D}_{\lambda} \tag{8}$$

对 $\boldsymbol{R}$ 的特征方程,有

$$\boldsymbol{R}\boldsymbol{\varphi}_i = \lambda_i \boldsymbol{\varphi}_i \tag{9}$$

式(9)中,λ_i 是自相关矩阵 $\boldsymbol{R}$ 的本征特征值,$\boldsymbol{\varphi}_i$ 是 λ_i 对应的本征向量。矩阵 $\boldsymbol{\varphi}$ 可由矩阵 $\boldsymbol{R}$ 的特征向量表示。

1.2　PCA 降维原理

将式(1)转换为下面的形式

$$\boldsymbol{x} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\varphi}_i + \sum_{i=m+1}^{n} \alpha_i \boldsymbol{\varphi}_i \tag{10}$$

通过去除后 $\boldsymbol{a}$ 的 $n-m$ 个特征向量时,则 $\boldsymbol{x}$ 的预估值是

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\varphi}_i \tag{11}$$

误差向量表示为:

$$\Delta\boldsymbol{x} = \boldsymbol{x} - \hat{\boldsymbol{x}} = \sum_{i=m+1}^{n} \alpha_i \boldsymbol{\varphi}_i \tag{12}$$

均方误差为:

$$\varepsilon^2 = \| \Delta\boldsymbol{x} \|^2 = \sum_{i=m+1}^{n} \alpha_i^2 = \sum_{i=m+1}^{n} \lambda_i \tag{13}$$

从式(13)可以看出,为达到均方误差最小的目的,只要使向量 $\boldsymbol{a}$ 对应的后 $n-m$ 个特征值之和最小即可。一般情况下,对向量 $\boldsymbol{a}$ 的特征值排序,变换矩阵 $\boldsymbol{\Phi}$ 由前 m 个特征值对应的特征向量组成。

$$\boldsymbol{\Phi}_m = (\varphi_1, \varphi_2, \cdots, \varphi_m) \tag{14}$$

利用 $\boldsymbol{\Phi}_m$ 代替 $\boldsymbol{\Phi}$,则有

$$\boldsymbol{a}_m = \boldsymbol{\Phi}_m^{\mathrm{T}}\boldsymbol{x} \tag{15}$$

向量 $\boldsymbol{a}_m$ 即为向量 $\boldsymbol{X}$ 降维后的主成分。经过 K-L 变换建立了一个新的坐标系,将原始向量映射到新的坐标系下。

2　传统的 Adaboost 算法

2.1　算法描述

Adaboost 算法的基本思想是针对同一个训练集选取单个特征训练分类器(弱分类器),对训练数据集进行 T 轮训练,将得到的弱分类器序列($h_1, \cdots, h_T$)经过其在训练集上的错误率加权构成一个更强的最终分类器(强分类器)。Adaboost 算法流程如表 1 所示。据分类结果和加权系数更新下次迭代的样本权重[第 3 步(e)],增加那些被错误分类样本的权重,并减少那些已经被正确分类的样本的权重。最后将训练得到的 T 个弱分类器组合得到强分类器。

2.2　算法分析

在传统的 Adaboost 算法中,第 3 步(a)弱分类器的训练,这个过程耗时最多,假设训练集有 k 个特征、N 个样本,训练弱分类器是在 k 个特征的简单分类器 h_j 寻找,每个 h_j 需搜索 N 个样本。假设训练一个 h_j 的时间为 time,样本 N 较大时,time 的值也较大。训练一个弱分类器 h_t 需找出 k 个 h_j 中的最小者,所需的时间是 $k \times time$。如果若干弱分类器组合成强分类器,训练时间会成倍的增多。研究者主要从两个角度对这个问题进行了研究,第一是对特征进行处理,第二是对样本进行处理。Wu 等[12]针对特征提出了两种快速训练方法:前向特征选择(forward feature selection,FFS)算法 Faster Adaboost 算法,其思路都是将和特征有关的每次迭代共享的计算放到迭代过程之前,通过共享的方式提高训练速度。贾慧星等[9]主要从样本方面进行研究,提出了动态权重裁剪的快速 Adaboost 训练算法,通过动态裁剪样本,缩短训练时间。严云洋等[13]主要从样本阈值搜索方式上进行改进,提高训练速度。

表 1　Adaboost 算法流程[9]

Table 1　Algorithmic process of Adaboost[9]

1)给定训练集 $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$ 和预定的迭代次数 T,其中 $\boldsymbol{x}_n \in X$ 是样本的特征矢量,$\boldsymbol{y}_n \in \{+1, -1\}$ 为样本的标签。
2)初始化样本权重 $d_n^{(1)} = 1/N, n = 1, 2, \cdots, N$。
3)迭代次数 $t = 1, 2, \cdots, T$。
(a)在当前的样本分布 $\{S, d^{(t)}\}$ 上训练弱分类器 $h_t: x \rightarrow \{+1, -1\}$;
(b)计算弱分类器 h_t 在当前样本分布 $d^{(t)}$ 上的加权错误率: $\varepsilon_t = \sum_{n=1}^{N} d_n^{(t)} I[y_n \neq h_t(x_n)]$;
(c)如果 $\varepsilon_t = 0$ 或者 $\varepsilon_t \geqslant 0.5$,则令 $T = t-1$,停止迭代;
(d)计算弱分类器 h_t 在最终分类器集合中的加权系数 $\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$;
(e)为下次迭代更新样本的权重 $d_n^{(t+1)} = d_n^{(t)} \exp[-\alpha_t y_n h_t(x_n)]/Z_t$,其中 Z_t 为归一化因子。
4)输出最后的强分类器: $H_T(x) = \mathrm{sign}[\sum_{t=1}^{T} \alpha_t h_t(x)]$。

3　Adaboost 算法改进

3.1　DRAdaBoost 分析与改进

传统的 Adaboost 算法是以每个样本的特征值作为预测阈值,虽然找点准确,但是泛化能力差。文

献[14]中提出 Rank-AdaBoost 算法,将每个特征对应的特征值空间 r 等份,

$$\Delta j = \{\max[(y_j(x)] - \min[y_j(x)]/r \quad (16)$$

式(16)中 $y_j(x)$ 表示第 j 个特征值。

$$\min[y_j(x)] + k\Delta j \quad (17)$$

再以式(17)的 r 个值作为阈值进行搜索,得到该特征的对应简单分类器。在很大程度上减少了求解简单分类器 $h_j(j = 1,\cdots,k)$ 的时间。

但该算法存在一个问题,没有考虑样本特征值空间维数。样本数量较多,针对每个特征的特征值个数可能比较少,若特征值空间维数为 V,$V < r$,对特征值空间进行 r 等份,相当于进行了 r-n 次无效求解。

针对上述算法存在的问题,提出了根据特征空间维数动态调整步进幅度的 DRAdaboost 算法。该算法对传统的 Adaboost 算法的第 3 步(a)进行了改进,算法的具体流程如下:

Step1 统计样本对应的第 j 个特征的特征空间维数 V_j;

Step2 判断 V_j 和 r 的大小,若 $V_j > r$,对特征空间进行 r 等份;若 $V_j < r$,对特征空间进行 V_j 等份。

3.2 基于 PCA 的算法改进

传统的 Adaboost 算法针对每个特征属性建立简单分类器,在简单分类器中搜索弱分类器。如果样本的特征属性较多时,搜索弱分类器是非常耗时的。样本的特征属性之间往往存在冗余,这会影响训练分类器的精度。针对上述问题,提出了 PCAdaboost 算法,利用 PCA 的降维技术减小样本的特征属性,降低属性间的相关性,提高训练精度。同时,在保证训练精度的情况下,由于减少了训练样本的特征属性个数,训练弱分类器的时间也会相应减少。具体的流程如下:

Step1 计算样本特征的协方差矩阵 $\boldsymbol{R}$;

Step2 计算协方差矩阵的特征值 λ 和特征向量 $\boldsymbol{\phi}$,选取使均方误差最小的特征值组合 λ_m,根据特征值组合获取降维矩阵 $\boldsymbol{\phi}_m$;

Step3 利用上一步获取的降维矩阵将样本映射到低维空间实现降维。

4 实验结果

为了比较算法的效果,实验选取了 UCI 数据中的 Horse,Hear 数据集进行分类实验,Horse 数据集有 3 个类标签,368 个样本,28 个特征值。Heart 数据集有 2 个类别标签,270 个样本,13 个特征值。弱分类器采用单层决策树(decision stump,也称决策树桩)。测试用的机器为 P340 2.20 GHz CPU,2 GB 内存,采用的 Python 版本是 2.7.6。

实验 1 观察 PCA 中主成分的个数与累计方差百分比的关系。通过图 1 可以看出,前 5 个主成分覆盖了 99% 的方差。这就表明了,如果保留前 5 个主成分,可以实现大概 3∶1的压缩比。

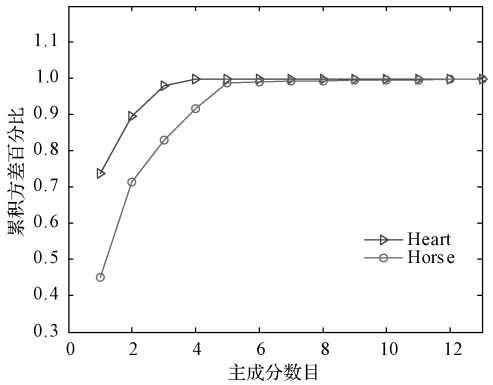

图 1 主成分个数与累计方差百分比的关系

Fig. 1 Relationship between number of principal components and cumulative percentage of variance

实验 2 不同的累积方差对训练错误率的影响。不同累积方差的选取没有太严格的步进间隔,由式(13)可知,不同累积方差的选取主要为特征值个数的选取。横坐标采用主成分个数来标定。Heart 数据集在前 4 个主成分就覆盖了数据 99.7% 的方差。从图 2 中可以看出只选取前 4 个主成分样本的训练错误率相对较高,降维过度可能损失有用信息。在以下的实验中针对 Heart 数据集选取前 8 个主成分作为样本特征集。Horse 数据集在前 6 个主成分覆盖数据 99.2%,增加主成分对训练错误率影响较小。对 Horse 数据集选取前 6 个主成分作为样本特征集。

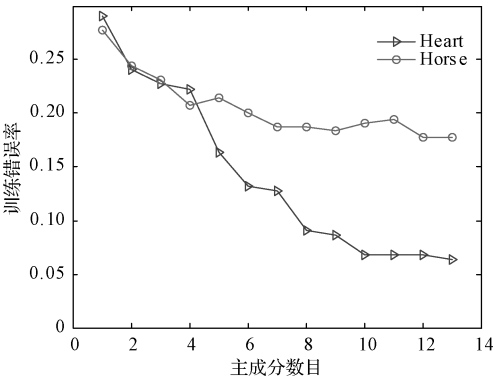

图 2 不同累积方差对训练错误率的影响

Fig. 2 Effect of training error rate for different cumulative variance

实验 3 不同改进算法的测试错误率比较。实验选取 Horse,Heart 数据集。作比较的算法为 adaboost、PCAdaboost、DRAdaboost、PCA + DRAdaboost。其中步进阈值选为 10,累积方法错误率选取包含

99% 信息量的主成分数量。依据实验 2 结果,Heart 和 Horse 数据集分别选取前 8 个和前 6 个主成分作为样本特征集。通过图 3 和图 4 可以看出,PCAdaboost 和 DRAdaboost 与原始的 adaboost 算法相比,在测试集上的错误率较小,PCA + DRAdaboost 在测试集上的错误率较原始 adaboost 算法相对较大。通过图 5 可以看出 PCAdaboost 在训练时间有提升,但提升幅度不大。DRAdaboost 和 PCA + DRAdaboost 在训练时间上的提升幅度较大。

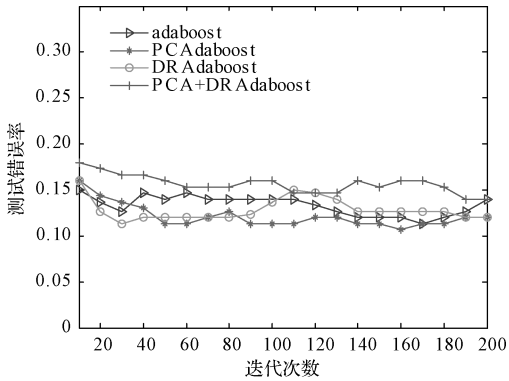

图 3　Horse 数据集的测试错误率比较

Fig. 3　Comparison of test error rate for Horse dataset

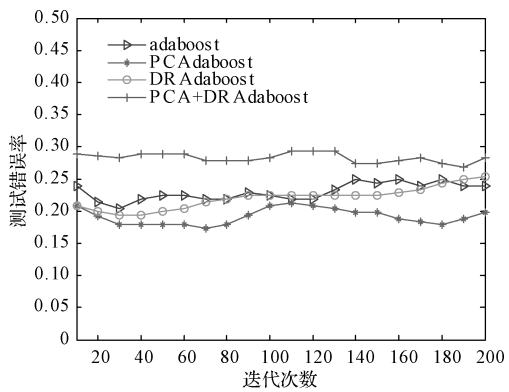

图 4　Heart 数据集的测试错误率比较

Fig. 4　Comparison of test error rate for Heart dataset

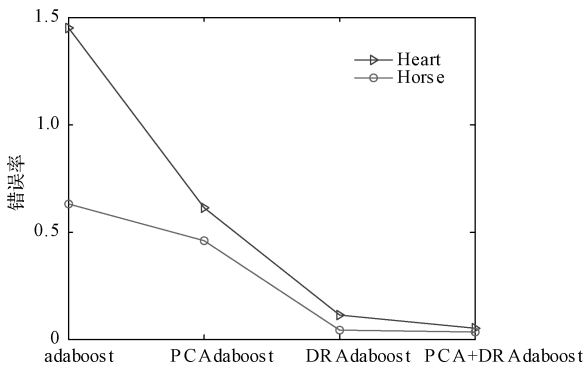

图 5　不同改进算法的训练速度比较

Fig. 5　Comparison of training speed for various modifications algorithm

由以上实验结果可知,PCAdaboost 和 DRAdaboost 算法的泛化能力较好,在测试集上的错误率较小。同时,DRAdaboost 和 PCA + DRAdaboost 在训练速度方面有很大提升。对于不同的训练数据集,如果样本特征属性较多,可利用 PCA 方法降维。样本量较多而特征较少,可利用 DRAdaboost 算法提高训练速度,尽管在训练精度上有欠缺,但是训练时间上的提升幅度很大。

5　结论

提出的基于 PCA 改进的快速 Adaboost 算法,它去除了样本属性间的相关性,保留了样本的隐含属性,在训练精度方面优于传统的 Adaboost 算法,同时在训练速度上有一定的提升。实验结果证明该算法在应对大规模数据时可以缩短训练时间,并能够保证训练性能,具有较大的实用价值。

参　考　文　献

1　Valiant L G. A theory of the learnable. Communications of ACM, 1984;27(11):1134—1142

2　Michael K, Valiant L G. Cryptographic limitations on learning Boolean formulae and finite automata. Journal of the ACM, 1994; 41 (1): 67—95

3　Schapire R E. The strength of weak learnability. Machine Learning, 1990; 5(2): 197—227

4　Freund Y. Boosting a weak algorithm by majority. Information and Computation, 1995; 121 (2): 256—285

5　Freund Y, Schapire R E. A decision -theoretic generation of online learning and an application to boosting. Journal of Computer and System Science, 1997;55(1):119—139

6　Paul V, Jones M J. Robust real-time face detection. International Journal of Computer Vision, 2004; 57(2):137—154

7　Sheng L, Ye X Q. Efficient lmprovement for adaboost based object detection. IEEE Computer Society,2009:95—98

8　范伯良,高　峰,寇　鹏. 在线 Boosting 回归算法及其在高耗能企业负荷预测中的应用. 信息与控制,2014;43(6):750—756
Fan B L, Gao F, Kou P. Online boosting regression method and its application to load forecasting in energy-intensive enterprise. Information and Control,2014;43(6):750—756

9　贾慧星,章毓晋. 基于动态权重裁剪的快速 Adaboost 训练算法. 计算机学报,2009;32(2):336—341
Jia H X, Zhang Y J. Fast adaboost training algorithm by dynamic weight trimming. Chinese Journal of Computers, 2009;32(2):336—341

10　赵　蕾,解争龙,李　红,等. 基于 PCA-*K*-means 的卫星遥感图像的颜色特征提取技术. 微电子学与计算机,2013;29(10):64—68
Zhao Q, Xie Z L, Li H, *et al*. Color-feature extraction of remote sensing image based on principal components analysis and *K*-means. Microelectronics & Computer,2013;29(10):64—68

11　侯小静. 基于 PCA 算法和人脸姿态合成的人脸识别. 长沙:中南大学,2013

Hou X J. Face recognition based on PCA algorithm and face pose synthesis. Changsha: Central South University, 2013

12 Wu J X, Brubaker S C, Mullin M D. Fast asymmetric learning for cascade face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008; 30(3): 369—382

13 严云洋,郭志波,杨静宇. 基于双阈值的增强型 AdaBoost 快速算法. 计算机工程,2007;33(21):172—174
Yan Y Y, Guo Z B, Yang J Y. Fast enhanced adaBoost algorithm based on dual-threshold. Computer Engineering, 2007; 33(21): 172—174

14 李文辉,倪洪印. 一种改进的 Adaboost 训练算法. 吉林大学学报(理学版),2011;49(3):498—504
Li W H, Ni H Y. An improved Adaboost training algorithm. Journal of Jilin University(Science Edition), 2011; 49(3): 498—504

Fast Adaboost Algorithm Based on Improved PCA

YUAN Shuang[1,2], LÜ Ci-xing[1]

(Shenyang Institute of Automation, Chinese Academy of Sciences[1], Shenyang 110016, P. R. China;
University of Chinese Academy of Sciences[2], Beijing 100049, P. R. China)

[**Abstract**] In view of the problem of the long training time in dealing with large training dataset in the training process of the traditional Adaboost algorithm, an improved methods was introduced to these problem. Improved algorithm using PCA dimension reduction technique, extracts main ingredients for the training sample feature, removes the correlation between the input sample characteristics, and improves the classification accuracy. At the same time, from the angle of sample threshold search takes into consideration the divisions and eigenvalue space dimension, threshold fast search method is presented. Experimental results show that the algorithm to achieve better results on UCI datasets.

[**Key words**] PCAdaboost principal components threshold search dimension reduction

(上接第 61 页)

11 Herout A, Dubská M, Havel J. Review of Hough transform for line detection. Real-time Detection of Lines and Grids, Springer London, 2013: 3—16

12 卢惠民. 机器人全向视觉系统自定位方法研究. 长沙:国防科学技术大学,2005
Lu Huimin. Research for robot self-localization system based on omnidirectional vision. Changsha: National University of Defense Technology, 2005

13 欧阳敏. RoboCup 中型组机器人定位与任意足球检测研究. 西安:长安大学, 2012
Ouyang Min. Research on the localization and arbitrary football detection of RoboCup middle-size robot. Xi′an: Chang'an University, 2012

Research of Self-localization System Based on Omnidirectional Vision for Soccer Robot

CAI Zong-yan[1], MA Peng-fei[1], OUYANG Min[2], DENG Xiao-kang[1]

(Key Laboratory for Highway Construction Technology and Equipment of Ministry of Education, Chang'an University[1], Xi'an 710064, P. R. China;
DongFeng Peugeot Citroen Automoble Company LTD Technology Center[2], Wuhan 430000, P. R. China)

[**Abstract**] In order to solve the self-localization problem in highly dynamic situation in the RoboCup middle size league competition, a fast robot self-localization method was used. Firstly, according to the structural of omnidirectional visual robot during visual processing, using an image mask the invalid information was removed. In order to avoid searching for all the pixels and narrow the scope of pixel, a method based on Bresenham algorithm was used. Using white line positioning algorithm to determine the robot position, combined with the search method based on Bresenham algorithm, the real-time image processing is effectively improved. Finally, a test was design based on ASRO Ⅱ robot and it is proved reliable and effective.

[**Key words**] omnidirectional visual information robot Bresenham algorithm scanning lines white line detection self-localization