文章编号：1673-9469（2019）01-0103-05 doi:10.3969/j.issn.1673-9469.2019.01.022

基于GA的负相关剪切集成不平衡行为分类研究

白梅娟，肖书忠，艾成伟，赵 超，黄 远，侯 帅，黄伟建

（河北工程大学 信息与电气工程学院，河北 邯郸 056038）

摘要：基于传感器的人类活动识别(HAR)在健康医疗领域具有重要的研究价值及研究意义。以往的关于传感器人类活动分类识别算法的研究，并没有考虑不同类别行为数据间的不平衡性。为了解决不同行为类别数据间的不平衡性影响算法精确度的问题，此算法采用下采样方法从大类和小类数据集中随机抽取选出若干组数量上相等的两种数据的集合，将多个不平衡数据变成平衡数据。其次，多个平衡数据集上训练多个弱分类器。然后，此算法以弱分类器的负相关和预测精度为代价函数，使用遗传算法挑选出能够使代价函数值最高的弱分类器来构成集成分类器 。使集成算法内的弱学习器具有较高预测精度和多样性。最后，此算法使用挑选出的弱学习器构成集成学习器对人的行为进行集成分类。此算法在已有的行为数据集上进行了仿真实验研究，实验结果证明本文提出的基于遗传的负相关剪切集成不平衡行为识别算法相对于传统算法能够有效提高不平衡行为识别的正确率。

关键词：行为识别；传感器；不平衡；剪切集成；多样性；遗传算法

中图分类号：TP391.4 **文献标识码**：A

Research on Negative Correlation Pruning Ensemble Imbalance Behavior Classification Based on Genetic Algorithm

BAI MEIjuan，XIAO Shuzhong，AI Chengwei，ZHAO Chao，HUANG Yuan，HOU Shuai，HUANG Weijian

(College of Information and Electrical Engineering，Hebei University of Engineering，Hebei Handan，056038，China)

Abstract：Sensor-based human activity recognition (HAR) has important research value and significance in the field of health care. Previous researches on classification and recognition of sensor human activities have not considered the imbalance between different categories of behavior data. In order to solve the problem that the imbalance between data of different behavioral categories affects the accuracy of the algorithm，our algorithm uses the downsampling method to randomly extract two sets of data from large and small data sets，which are equal in number，and transform multiple imbalanced data into balanced data. Secondly，multiple weak classifiers are trained on multiple balanced datasets. Then，the algorithm takes the negative correlation and prediction accuracy of the weak classifier as the cost function and uses genetic algorithm to select the weak classifier which can make the highest value of the cost function to form the integrated classifier. The weak learner in the ensemble algorithm has high prediction accuracy and diversity. Finally，the algorithm uses the selected weak learners to construct an ensemble learner to classify human behavior. The experimental results show that the proposed algorithm can effectively improve the accuracy of unbalanced behavior recognition compared with the traditional algorithms.

Key words：behavior recognition；sensor；imbalance；pruning ensemble；diversity；genetic algorithm

收稿日期：2018-12-02 特约专稿

基金项目：河北省自然科学基金资助项目(E2017402115) 河北省自然科学基金资助项目(E2015402077)；河北省教育厅高校科学技术研究青年基金资助项目(QN2018073)

作者简介：白梅娟，(1990-)，女，河北邯郸人，硕士，助理实验师，从事模式识别与图像处理方面的研究。

基于传感器的人体行为识别技术具有低功耗，高便携性，高隐私的特点，在老人监护、重病患者监护和儿童监护等健康医疗方面具有重要的研究价值。相对于基于视频的活动识别系统，基于传感器的人类活动识别系统由加速度计、陀螺仪等行为识别传感器构成。

以往的关于传感器人类行为识别的算法研究，忽略了不同行为类别间的不平衡特性，导致算法难以精确识别人体的少类行为。已有的关于剪切集成算法中的弱学习器的挑选过程没有考虑不同弱学习器在不同类别不平衡数据集上的多样性问题，降低了剪切集成算法的泛化性能。为解决人体行为识别的不平衡问题和提高弱学习器的多样性，本文采用下采样方法从大类和小类数据集中随机抽取选出若干组数量上相等的两种数据的集合，将多个不平衡数据变成平衡数据。本文提出了一种基于遗传算法的负相关剪切集成不平衡行为识别算法。

1 基于遗传的负相关剪切集成不平衡行为识别算法

1.1 基于传感器的行为识别算法研究现状

不平衡问题是数据挖掘中十大最有挑战性的问题之一。类别不平衡数据是指其中一个类别数量远超过其他类的数量。为了保证模型的预测精度，模型训练时倾向于将小类分错分为大类样本，以提高训练模型的预测精度。

国内外许多高校和研究机构以及学者对基于传感器的行为识别算法进行了大量的研究 [1-15]。

1.2 本文算法模型与框架

以往的基于传感器的人体活动识别算法忽略了数据的不平衡特性，导致算法难以高精度地识别人体的少数类行为，但是往往一些少数类行为是我们更加关注的，例如老人跌倒等动作往往属于少数类动作，但是这些动作往往是我们在老人监护中更期望关注的行为。为解决人体行为识别的不平衡问题，本文采用下采样方法从大类和小类数据集中随机抽取选出若干组数量上相等的两种数据的集合，将多个不平衡数据变成平衡数据。其次，多个平衡数据集上训练多个弱分类器。然后，本文以弱分类器的负相关和预测精度为代价函数，使用遗传算法挑选出能够使代价函数值最高的弱分类器来构成集成分类器 。使集成算法内的弱学习器具有较高预测精度和多样性。最后，本文使用挑选出的弱学习器构成集成学习器对人的行为进行集成分类。

集成学习中弱分类器的预测精度和弱分类器的多样性是影响集成学习的两个重要指标，以往的研究仅仅考虑数据集的整体多样性，并没有分别考虑多类和小类间的多样性。本论文充分考虑了分类器在多数类别和小类上的多样性 [16]。下面公式 (1) 给出了本文的代价函数：

$$e_i = \frac{1}{2}\left(h_i\left(x_p\right) - y_p\right)^2 - \lambda\left(h_i\left(x_p\right) - H\left(x_p\right)\right) \times \sum_{j\neq 1}\left(h_i\left(x_p\right)\right) - H\left(x_p\right)\right) \tag{1}$$

$$\lambda\begin{cases}\lambda_{min}, & \text{if } x \in \text{小类} \\ \lambda_{max}, & \text{if } x \in \text{大类}\end{cases} \tag{2}$$

公式 (1) 中 x_p 代表被训练数据，y_p 代表数据的真实值，$h_i(x_p)$ 表示弱分类器对数据的预测值，$H(x_p)$ 表示集成的分类器。$\lambda \in (0,1)$ 表示分类器的学习参数，分为两种，一种是对大类数据的学习权值 λ_{max}，另外一种是对小类数据的学习权值 λ_{min}。公式 (2) 中 $\lambda_{\min} > \lambda_{\max}$。式 (1) 中第一项 $(h_i(x_p) - y_p)^2$ 为分类器预测值与真实值之间的误差，第 2 项 $\lambda\left(h_i\left(x_p\right) - H\left(x_p\right)\right) \times \sum_{j\neq 1}\left(h_i\left(x_p\right) - H\left(x_p\right)\right)$ 为矫正惩罚函数。第二项的目的是使当前学习器与集成学习器中剩余的学习器之间负相关。在训练的过程中，每一个弱学习器不仅仅只是减小自己预测值和真实值之间的误差，同时也会增大弱学习器之间的差异性，从而提高小类的多样性。

算法步骤

步骤 1：初始化

训练集 $S=\{x_p, y_p\}$，分类器集合 $H=\{h_i, \cdots h_k\}$，n 为弱学习器个数，t 为遗传算法最终挑选出来

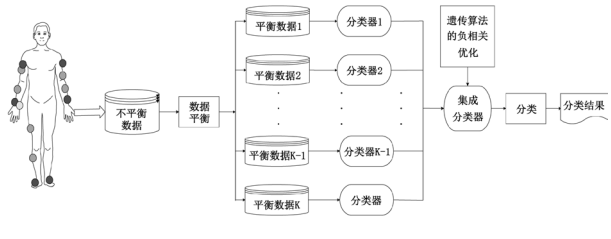

图 1 基于遗传的负相关剪切集成不平衡行为识别算法示意图

Fig.1 The diagram of behavior identification algorithm of negative imbalance pruning ensemble handling imbalance based on genetic algorithm

的弱学习器个数。Gen 为遗传算法迭代的次数。

步骤 2：

对行为数据集采用underbagging算法进行建模，建立行为识别的 underbagging 预测模型。模型内的弱学习器为 hi，强分类器为 H。

步骤 3：

For m=1：*Gen*

算法以公式 (1) 为遗传算法的 *fitness* 函数，选择出 *gen* 个弱学习器，使 *m* 个弱学习器集成后的 *AUC* 为最高。

End

步骤 4：

对测试集的样本进行分类，并输出预测精度。

1.3 UnderBagging 集成算法

UnderBagging 算法对数据集进行随机抽取，抽取出不同的训练样本集合，然后不同的弱分类器在不同的训练样本集合上进行训练，训练出的弱分类器构成一个强分类器 [17]。强分类器对未分类的预测样本进行预测分类。集成算法中的弱学习器精度和弱分类器多样性是影响集成算法预测精度的两个关键指标 [17]。Underbagging 算法对不平衡问题进行了处理，但是没有挑选弱分类器。

1.4 负相关学习

负相关学习 (Negative correlation learning，NCL) 是一种常用于神经网络集成的集成学习算法，它是把基学习器之间的差异性作为一个显性的度量标准引入到神经网络的损失函数中去，进而影响神经网络的训练。通过调整影响因子可以权衡基神经网络之间的性能与多样性，以谋求获得一个性能最优的集成神经网络模型。

NCL 算法创造出不同的弱学习器，让弱学习器学习数据的不同部分，提高学习器之间的多样性 [18-19]，以下式为例。假设数据集 T={(x_p，y_p，p=1，…，…，N)}，$H(x_p)$ 为强学习器，其定义如下：

$$H(x_p)=\frac{1}{M}\sum_{i=1}^{M}h_i(x_p) \tag{3}$$

式中，M 为弱学习器分类个数，$h_i(x_p)$ 为弱学习器对 x_p 样本的预测值，$H(x_p)$ 为集成输出。

C_i 为集成学习的负相关值：

$$C_i=\left(h_i(x_p)-H\left(x_p\right)\right)\times\sum_{j\neq 1}\left(\left(x_p\right)-H(x_p)\right)=-\left(h_i(x_p)-H(x_p)\right)^2 \tag{4}$$

NCL 算法提高集成学习内不同弱学习器间的多样性。然而以往的关于集成算法 NCL 多样性的研究往往关注的是样本的整体多样性，并没有关注不同类别内弱学习器的多样性。

1.5 遗传优化算法

达尔文的“物竞天择，适者生存”生物进化论是遗传优化算法的基础理论。遗传优化算法是生物学科和数学学科的交叉学科领域的一种寻找最优解的优化算法 [20]。遗传算法适用于待解的问题为一个集合，需要寻找最优解的情景。通过模仿人体的基因编码的遗传学机制，遗传算法给待解的集合进行二进制编码，成为初代的种群。初代的种群经过基因的交叉和变异，产生二代和三代，按照优胜略汰的原理，每一代中会淘汰一部分不符合限制条件的个体，直到算法收敛，末代种群中的个体作为待解问题的最优解 [20]。

本文算法采用遗传优化算法，遗传优化算法的代价函数是以弱分类器个体的预测精度和对不同类别数据识别的弱分类器的多样性为基础建立的。遗传算法收敛后，从弱分类器集合中，选出最优的弱分类器集合作为集成分类器，进行数据集的预测和分类。

2 实验仿真

2.1 实验仿真环境和数据集

本文算法在 OPPORTUNITY 活动识别数据集、Daily and Sports Activities Data Set 和 Human Activity Recognition Using Smartphones Dataset 3 个已有的数据集上进行了仿真实验。实验仿真平台：MATLAB 2017B。

从机器学习 UCI 数据库中选取的 OPPORTUNITY 人体行为识别数据集的原始数据由人体穿戴传感器、环境传感器和目标传感器等三种传感器采集得到，传感器类型包含加速度计 (Acc)，陀螺仪 (Gyro)，磁力传感器 (Magn)，四元数 (Quat)，霍尔传感器，航向传感器，身体传感器，身体坐标系角度旋转速度传感器和地坐标系角度旋转速度传感器。在本文实验部分使用的数据为从人体上所携带的传感器网络采集得出，数据集中的前 96 个属性由人体腰部以上的 5 个惯性传感器单元与鞋子上的 2 个惯性传感器单元采集得到，

随后的 36 个属性由放置在人的上肢、臀部和腿等部位的 3 轴加速传感器单元采集，剩余的 12 个属性由放置在肩部左 / 右前 / 后侧的超宽带定位系统记录。OPPORTUNITY 包含有四种行为模式：站立 (A1)，走 (A2)，坐 (A3) 和躺 (A4)。所有活动由 4 名受试者通过每日 5 次重复跑步进行。Daily and Sports Activities Data Set：该数据集是日常运动行为数据集。该数据集合由 5 个运动传感器构成。该数据集由年龄在 20 ～ 30 周岁间的 4 个女性和 4 个男性。数据集内包含有 19 个行为，本文以 sit-ting (A1) 和 standing (A2) 为行为识别目标进行行为识别。Human Activity Recognition Using Smartphones Dataset 是智能手机人类行为识别数据集。该数据集是来源于对年龄 19-48 年龄段内的 30 个志愿者来每个人佩戴的智能三星手机 Samsung Galaxy S II 进行的 6 个行为 (走，上楼梯，下楼梯，坐，站，躺) 数据集。本文对走和躺两种行为进行识别。

为了防止不平衡数据集的数据偏移现象，本论文采用了 DOB-SCV 对数据集合进行了提取。

在数据的采集过程中，传感器网络中的无线传感器都可能由于连接不稳定而断开，造成一部分数据丢失出现空值。在特征提取之前，首先需将数据集进行清洗和信号分割。文中使用两种方式来处理丢失的数据。第一种方式将丢失一半以上数据的传感器对应的属性全部删除。第二种方式是传感器的缺失数据少于一半时，简单地重复同一维度上的值来代替缺失值。由于原始数据采集时进行窗口分割，由此本文使用长度为 500 ms，步长为 250 ms 的滑动窗口来进行样本点的划分。

2.2 实验结果

本文采用公认的 (Area Under the receiver operating characteristic Curve)AUC 参数衡量行为识别算法的预测精度。AUC 值是一种对分类器的平均预测精度衡量的参数。下面公式 (5) 给出了分类器的 AUC 的值的计算方法：

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{5}$$

公式 (5) 中，TP_{rate} 表示分类器把正样本预测为正样本的概率，FP_{rate} 表示分类器把负样本预测为正样本的概率。

实验运行后，实验的结果如下图 2 所示：

从图 2 中可以看出，当弱学习器个数为 8 的时候，挑选出来的弱学习器构成的集成分类器算法预测效果最好，而将所有的弱学习器集成为一个强学习器后的预测效果并不是最优的。

本文算法和 3 种基于传感器的行为识别算法进行了预测精度比较，作比较的三种算法分别为：Bagging、Underbagging 和 Negative correlation 算法。每种算法的预测精度都是用 AUC 值来衡量。四种算法的预测 AUC 值比较结果如表 1。

表 1 中，每个算法的精度是以 AUC 值来计算的，实验结果加上了随机误差。从表 1 中可以看出，对三个人体行为数据集的预测精度，本文提出的算法预测精度均最高。本文相对于其他算法的预测精度较高的原因如下：Bagging 算法是一种普通的集成

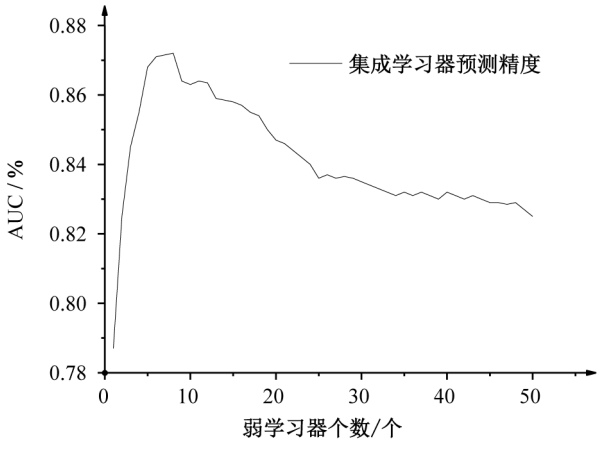

图 2 不同数量弱学习器下集成算法的预测精度图

Fig.2 The Prediction accuracy diagram of ensemble algorithm under different number of weak learning devices

表 1 不同算法的行为识别精度比较表

Tab.1 Table of comparison of behavior recognition accuracy of different algorithms

算法 \ 数据集	Opportunity	Daily and Sports Activities Data Set	Human Activity Recognition Using Smartphones Dataset
Bagging	0.821 2 ± 0.006 3	0.877 5 ± 0.010 0	0.827 5 ± 0.007 2
Underbagging	0.821 2 ± 0.006 8	0.887 5 ± 0.010 0	0.847 5 ± 0.018 4
Negative correlation	0.821 2 ± 0.018 2	0.889 4 ± 0.014 9	0.859 4 ± 0.011 2
The proposed algorithm	0.849 3 ± 0.010 6	0.890 1 ± 0.006 3	0.860 1 ± 0.038 8

算法，没有对不平衡问题进行处理，并且该算法没有对弱学习器进行筛选，集成分类器内部可能存在恶化性能的弱学习器。Underbagging 算法对不平衡问题进行了处理，但是没有挑选弱分类器，集成分类器内部依然可能会存在一些会降低集成算法预测精度的弱学习器。Negative correlation 算法对弱分类器进行了挑选，但是该算法并没有考虑数据集大类和小类数据的不平衡性，没有考虑多类和小类间的多样性差异，仅仅考虑了数据集整体的多样性。本文的算法不仅考虑了数据集内大类和小类样本的多样性问题，还考虑了数据大类和小类之间的不平衡性，而且对集成算法中的弱学习器采用遗传优化算法进行了筛选，去除了恶化集成分类器性能的弱分类器，所以本文算法精确度较高。

3 结论

1) 本文提出了一种基于遗传算法的负相关剪切集成不平衡行为识别算法，本文算法在剪切集成算法挑选弱学习器时，兼顾了弱学习器在行为大类样本和小类样本上的多样性。能够有效的提高算法的稳定性。

2) 本文采用遗传优化算法对弱分类器进行挑选，将精度高和多样性高的弱学习器从集成算法中挑选出来，构成了一个强分类器，提高了集成分类器的预测精度。本文提出的算法与 3 个传统的算法进行了比较，实验结果表明本文算法具有更高的识别精度。

参考文献：

[1] 衡 霞，王忠民 . 基于手机加速度传感器的人体行为识别 [J]. 西安邮电大学学报，2014，19(6)：76-79.

[2] 徐炳雪，史建华，钱俊臣，等 . 基于加速度传感器的人体行为识别系统的设计与实现 [J]. 电脑开发与应用，2014，27(12)：55-57.

[3] 高 蕾，曹建忠 . 基于可穿戴传感器的行为识别随机逼近模型 [J]. 计算机技术与发展，2014，24(12)：83-87.

[4] 杨璐璐，陈建新，周 亮，等 . 基于无线体域网的囚犯异常行为实时分析 [J]. 计算机科学，2015，42(03)：47-50.

[5] 段梦琴，李仁发，黄 晶 . 融合关联性的多任务压缩感知行为识别方法 [J]. 计算机工程与科学，2015，37(06)：1071-1078.

[6] 周博翔，李 平，李 莲 . 改进随机森林及其在人体姿态识别中的应用 [J]. 计算机工程与应用，2015，51(16)：86-92.

[7] 罗 坚，唐 琏，毛 芳，等 . 基于云计算的可穿戴式老龄人异常行为检测系研究 [J]. 传感技术学报，2015，28(8)：1108-1114.

[8] 何 鹏，陈跃跃，扈 啸 . 基于智能手表加速度传感器的人体行为识别 [J]. 电脑与信息技术，2015，23(5)：6-8.

[9] 卢先领，王洪斌，王莹莹，等 . 加速度数据特征在人体行为识别中的应用研究 [J]. 计算机工程，2014，40(5)：178-182.

[10] 郭东东，郝润芳，吉增涛，等 . 基于三轴加速度传感器的山羊行为特征分类与识别 [J]. 家畜生态学报，2014，35(8)：53-57.

[11] 张 洁 . 基于加速度传感器的人体运动行为识别研究 [J]. 自动化与仪器仪表，2016，(3)：228-229.

[12] 余 杰 . 基于加速度传感器的人体行为识别研究 [D]. 天津 天津工业大学，2017.

[13] 强茂山，张东成，江汉臣 . 基于加速度传感器的建筑工人施工行为识别方法 [J]. 清华大学学报 (自然科学版)，2017，57(12)：1338-1344.

[14] 郑增威，杜俊杰，霍梅梅，等 . 基于可穿戴传感器的人体活动识别研究综述 [J]. 计算机应用，2018，38(5)：1223-1229.

[15] 朱响斌，邱慧玲 . 基于智能手机传感器数据的人类行为识别研究 [J]. 计算机工程与应用，2016，52(23)：1-5.

[16]SUN Z，SONG Q，ZHU X，et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition，2015，48(5)：1623-1637.

[17]HOU S，HUA F，LV W，et al. Hybrid Modeling of Flotation Height in Air Flotation Oven Based on Selective Bagging Ensemble Method[J]. Mathematical Problems in Engineering，2013，2013(3)：1-9.

[18]WANG S，TANG K，YAO X. Diversity exploration and negative correlation learning on imbalanced data sets[C]. International Joint Conference on Neural Networks，2009：1796-1803.

[19]LIU Y，YAO X. Simultaneous training of negatively correlated neural networks in an ensemble[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society，1999，29(6)：716.

[20]MADRASWALA H S，DESHPANDE A S. Genetic algorithm solution to unit commitment problem[C]. International Conference on Nascent Technologies in Engineering，2017：1-6.

(责任编辑　李新)