(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110458204 A
(43)申请公布日 2019.11.15

(21)申请号 201910664303.6

(22)申请日 2019.07.23

(71)申请人 上海交通大学
地址 200240 上海市闵行区东川路800号

(72)发明人 朱平 颜诗旋 刘钊 刘灿

(74)专利代理机构 上海交达专利事务所 31201
代理人 王毓理 王锡麟

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 20/00(2019.01)

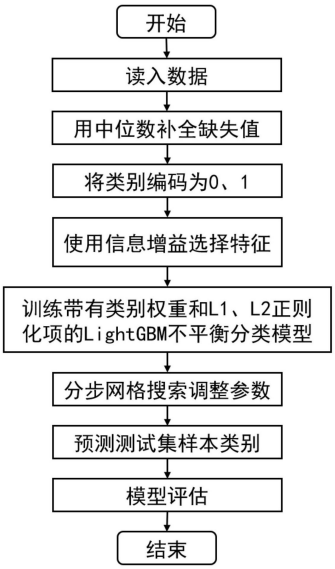
权利要求书1页 说明书4页 附图2页

(54)发明名称

基于信息增益和LightGBM模型的汽车故障预测方法

(57)摘要

一种基于信息增益和LightGBM模型的汽车故障预测方法,以信息增益值作为评价指标度量特征与类别间的相关程度进行特征选择和训练样本的生成,通过训练样本对LightGBM不平衡模型训练后,进一步采用分步网格搜索优化模型参数并将优化后的模型用于汽车故障预测;本发明提升模型效率的同时提高了故障查全率,从而显著增强了对汽车故障的预测能力。



CN 110458204 A

1.一种基于信息增益和LightGBM模型的汽车故障预测方法，其特征在于，以信息增益值作为评价指标度量特征与类别间的相关程度进行特征选择和训练样本的生成，通过训练样本对LightGBM不平衡模型训练后，进一步采用分步网格搜索优化模型参数并将优化后的模型用于汽车故障预测；

所述的LightGBM不平衡分类模型是指：以决策树为基学习器的集成学习模型，通过使用直方图算法寻找决策树的最佳分裂结点，并使用带深度限制的叶子生长策略分裂结点，该模型在损失函数中引入了类别权重和L_1、L_2正则化项，具体为：修正损失函数$L(\phi)=\sum_i a_i\cdot l(\hat{y}_i,y_i)+\alpha\|\omega\|_1+\beta\|\omega\|_2^2$，其中：类别权重系数$a_i=\begin{cases}1,if\ y_i=0\\ \lambda,if\ y_i=1\end{cases}$，出于放大少数类损失的目的，将少数类权重系数γ设置为一个大于1的整数，$l(\hat{y}_i,y_i)$是单棵决策树对样本类别y_i和预测类别$\hat{y}_i$的损失函数，α||ω||$_1$为L_1正则化项，$\beta\|\omega\|_2^2$为L_2正则化项，ω为决策树的参数，在模型训练过程将由决策树算法自动设定，α、β为正则化项系数。

2.根据权利要求1所述的汽车故障预测方法，其特征是，所述的信息增益是指：某特征所提供的类别可分性的信息，定义为先验熵H(F)与后验熵H(F|Y)的差值：IG(F;Y)＝H(F)-H(F|Y)，其中：特征F的先验熵H(F)＝$-\sum_i P(f_i)\log_2(P(f_i))$，其中：P(f)为特征f的概率密度函数；特征F对类别Y的后验熵H(F|Y)＝$-\sum_j P(y_j)\sum_i P(f_i|y_j)\log_2(P(f_i|y_j))$，其中：P(f|y)为特征f对类别Y的条件概率密度函数；

在计算出各特征的信息增益后，对各特征的信息增益按照从大到小进行降序排列，从而剔除掉排名靠后的20个特征，使用余下的特征送入模型进行训练。

3.根据权利要求1所述的汽车故障预测方法，其特征是，所述的LightGBM不平衡分类模型训练是指：模型的损失函数最小化的过程，对样本数量为m、特征维度为n的数据集D＝{(x_i,y_i)}，其中：x_i为第i个样本，y_i为该样本的类别，y_i＝0为无故障，y_i＝1为有故障，在训练集上使用LightGBM不平衡分类模型训练时，以损失函数最小为目标进行迭代。

4.根据权利要求1所述的汽车故障预测方法，其特征是，所述的分步网格搜索是指：先使用较广的搜索范围和较大的步长，寻找全局最优值，即L_1正则化项系数α、L_2正则化项系数β、少数类权重系数γ可能的位置，然后逐渐缩小搜索范围和步长，来寻找更精确的最优值。

5.根据权利要求1所述的汽车故障预测方法，其特征是，在测试集上使用查全率评价预测性能，该查全率是指故障被机器模型能够成功预测到的概率，即$Recall=\frac{TP}{TP+FN}$，其反映了模型对汽车故障样本的预测能力，其中：TP为被正确分类为有故障的样本数，即有故障的样本被成功预测为有故障；FP为被错误分类为有故障的样本数，即无故障的样本被误认为有故障；TN为被正确分类为无故障的样本数，即无故障的样本被成功预测为无故障；FN为被错误分类为无故障的样本数，即有故障的样本被误认为无故障。

基于信息增益和LightGBM模型的汽车故障预测方法

技术领域

[0001] 本发明涉及的是一种汽车制造领域的技术，具体是一种基于信息增益和LightGBM模型的汽车故障预测方法。

背景技术

[0002] 汽车故障预测是指对于收集到的汽车故障数据集，建立机器学习模型，从而预测新的样本所属的类别，即故障或正常，从而对有故障的汽车及时进行检修，将汽车故障引起的交通事故防患于未然。

[0003] 收集到的汽车故障数据集常呈现出特征维度高、类别不平衡的特点，而现有的汽车故障预测方法大多忽视了这两个特点，导致故障查全率较低。如何准确地量化特征与与类别的相关性以剔除掉不相关的特征，并增强对类别不平衡数据集的预测能力，是汽车故障预测中亟待解决的问题。

发明内容

[0004] 本发明针对现有方法的不足，提出一种基于信息增益和LightGBM模型的汽车故障预测方法，使用信息增益衡量特征与与类别的相关性，进而剔除了不相关的特征；针对类别不平衡问题，建立了带有类别权重和L1、L2正则化项的LightGBM不平衡分类模型，提高了对故障的查全率。

[0005] 本发明是通过以下技术方案实现的：

[0006] 本发明涉及一种基于信息增益和LightGBM模型的汽车故障预测方法，以信息增益值作为评价指标度量特征与类别间的相关程度进行特征选择和训练样本的生成，使用训练样本对LightGBM不平衡分类模型训练后，进一步采用分步网格搜索优化模型参数并将优化后的模型用于汽车故障预测。

[0007] 所述的信息增益(Information Gain，IG)是指：某特征所提供的类别可分性的信息，定义为先验熵H(F)与后验熵H(F|Y)的差值：IG(F;Y)＝H(F)-H(F|Y)，其中：特征F的先验熵$H(F)=-\sum_i P(f_i)\log_2(P(f_i))$，其中：P(f)为特征f的概率密度函数；特征F对类别Y的后验熵$H(F|Y)=-\sum_j P(y_j)\sum_i P(f_i|y_j)\log_2(P(f_i|y_j))$，其中：P(f|y)为特征f对类别Y的条件概率密度函数。

[0008] 所述的特征是指：样本在某方面的性质，包括但不限于汽车的速度、行驶里程等。

[0009] 所述的类别是指：样本所属的类别，在汽车故障预测中类别为发生故障或状态正常。

[0010] 所述的特征选择是指：计算出各特征的信息增益后，对各特征的信息增益按照从大到小进行降序排列，从而剔除掉排名靠后的特征，使用余下的特征送入模型进行训练。

[0011] 所述的LightGBM不平衡分类模型是指：以决策树为基学习器的集成学习模型，通过使用直方图算法寻找决策树的最佳分裂结点，并使用带深度限制的叶子生长策略分裂结点，该模型在损失函数中引入了类别权重和L_1、L_2正则化项，具体为：修正损失函数$L(\phi)=$

$\sum_i a_i \cdot l(\hat{y}_i, y_i) + \alpha\|\omega\|_1 + \beta\|\omega\|_2^2$，其中：类别权重系数$a_i = \begin{cases} 1, if\ y_i = 0 \\ \lambda, if\ y_i = 1 \end{cases}$，出于放大少数类损失的目的，将少数类权重系数γ设置为一个大于1的整数，初始状态下设置为10，$l(\hat{y}_i, y_i)$是单棵决策树对样本类别y_i和预测类别$\hat{y}_i$的损失函数，α||ω||$_1$为L_1正则化项，$\beta\|\omega\|_2^2$为L_2正则化项，ω为决策树的参数，在模型训练过程将由决策树算法自动设定，α、β为正则化项系数，初始状态下均设置为0.1。

[0012] 所述的损失函数是指在模型训练过程中量化模型的预测类别与真实类别之间的差异的函数。对于标准的LightGBM模型，其损失函数为：$L(\phi) = \sum_i l(\hat{y}_i, y_i)$，$y_i$为该样本的类别，$\hat{y}_i$为单棵决策树对第i个样本的预测类别，$l(\hat{y}_i, y_i)$是单棵决策树对样本类别$y_i$和预测类别$\hat{y}_i$的损失函数。

[0013] 所述的LightGBM不平衡分类模型中的类别权重是为数据集中的少数类(有故障)样本、多数类(无故障)样本设置不同的重要性，使得少数类样本在模型训练过程中更为重要，达到放大少数类样本损失的目的，加强对少数类的学习。

[0014] 所述的LightGBM不平衡分类模型中的L_1正则化倾向于使得模型参数尽量稀疏，即非零分量个数尽量少，L_2正则化倾向于使得模型参数尽量均衡，即非0参数个数尽量稠密。为了避免仅适用L_1正则化使模型参数过于系数或仅适用L_2正则化使模型参数过于稠密，因而在损失函数中同时引入L_1、L_2两个正则化项，以有效地控制模型复杂程度。

[0015] 所述的LightGBM不平衡分类模型训练是指：模型的损失函数最小化的过程。对样本数量为m、特征维度为n的数据集D＝{(x_i,y_i)}，其中：x_i为第i个样本，y_i为该样本的类别，y_i＝0为多数类(无故障)，y_i＝1为少数类(有故障)，在训练集上使用本发明的LightGBM模型训练时，以损失函数最小为目标进行迭代。

[0016] 所述的分步网格搜索是指：先使用较广的搜索范围和较大的步长，寻找全局最优值，即L_1正则化项系数α、L_2正则化项系数β、少数类权重系数γ可能的位置，然后逐渐缩小搜索范围和步长，来寻找更精确的最优值。

[0017] 本发明进一步优选在测试集上使用查全率评价预测性能，该查全率是指故障被机器模型能够成功预测到的概率，即$Recall = \frac{TP}{TP+FN}$，其反映了模型对汽车故障样本的预测能力，其中：TP为被正确分类为有故障的样本数，即有故障的样本被成功预测为有故障；FP为被错误分类为有故障的样本数，即无故障的样本被误认为有故障；TN为被正确分类为无故障的样本数，即无故障的样本被成功预测为无故障；FN为被错误分类为无故障的样本数，即有故障的样本被误认为无故障。

技术效果

[0018] 与现有技术相比，本发明使用信息增益评价特征与类别间的相关性，有效降低了特征维度；本发明建立了带有类别权重和L1、L2正则化项的LightGBM不平衡分类模型，并使用分步网格搜索给出参数的最优取值，提升模型效率的同时提高了故障查全率，从而增强了对汽车故障的预测能力。

附图说明

[0019] 图1为本发明流程示意图；

[0020] 图2为实施例中缺失值比例最高的20个特征的柱状图；

[0021] 图3为实施例中信息增益最小的20个特征的柱状图。

具体实施方式

[0022] 如图1所示，本实施例以斯堪尼亚卡车汽车故障预测数据集为例进行说明，具体包括以下步骤：

[0023] 步骤1、读取数据：本实施例所采用的数据集特征维度为170维，记录了汽车速度、行驶里程、档位等信息。数据集共有60000个训练样本和16000个测试样本，各样本类别为有故障或无故障。其中在训练集的60000个样本中，有59000个样本的类别为无故障，仅有1000个样本的类别为有故障。

[0024] 步骤2、用中位数补全缺失值：统计数据集中各特征的缺失值比例，缺失值比例最高的20个特征如图2所示，可见数据集中存在大量的缺失值，本发明使用各特征的中位数补全缺失值。

[0025] 步骤3、类别编码：将类别编码为0、1，将无故障样本的类别编码为0，将有故障样本的类别编码为1。

[0026] 步骤4、使用信息增益选择特征：使用信息增益IG统计各特征的重要程度，信息增益最小的20个特征如图3所示。考虑到信息增益小的特征所提供的类别可分性信息较少，本实施例中剔除掉信息增益最小的20个特征，使用余下的150个特征作为模型的训练样本。

[0027] 步骤5、对带有类别权重和L1、L2正则化项的LightGBM不平衡分类模型进行训练：在训练集上，使用本方法所提出的带有类别权重和L1、L2正则化项的LightGBM不平衡分类模型作为学习器，按照5折交叉验证的方式训练模型。

[0028] 步骤6、分步网格搜索优化：使用分步网格搜索的方法调整LightGBM不平衡分类模型的L_1正则化项系数α、L_2正则化项系数β、少数类权重系数γ等3个参数。

[0029] 本实施例经上述优化后得到的参数为：α＝0.01，β＝0.005，γ＝55。

[0030] 步骤7、预测测试集样本类别：在测试集上使用训练好的LightGBM机器学习模型，预测16000个测试集样本的类别。

[0031] 本实施例中对每个样本的分类阈值设置为0.05。

[0032] 步骤8、模型评估：在测试集上通过混淆矩阵计算查全率。

[0033] 表1混淆矩阵

	实际有故障	实际无故障
预测有故障	365	679
预测无故障	10	14946

[0034] 根据表1可见，测试集中实际上共有375个样本的类别为有故障，本方法从中成功预测出了365个故障样本，对故障样本的查全率达97.33％，很好地完成了对汽车故障的预测。

[0035] 为了进一步说明本方法的有效性，使用传统的GBM机器学习方法作为对比，分别计算各自的故障查全率Recall，并记录各方法的模型训练时间，以评价各方法的预测性能。本方法所得结果与GBM机器学习方法比较如表2所示。

[0036] 表2本方法与GBM方法的结果对比

	故障查全率	模型训练时间
本发明	97.3%	13 s
GBM	67.2%	284 s

[0037] 根据表2可见，本方法很好实现了预期的发明目的，故障查全率达97.3％，相比传统的GBM方法提高30.1％，增强了对汽车故障的预测能力；另一方面，本方法相比传统的GBM方法，大幅度降低了模型训练时长，提高了使用机器学习技术预测汽车故障的建模效率，降低了计算成本。

[0038] 上述具体实施可由本领域技术人员在不背离本发明原理和宗旨的前提下以不同的方式对其进行局部调整，本发明的保护范围以权利要求书为准且不由上述具体实施所限，在其范围内的各个实现方案均受本发明之约束。

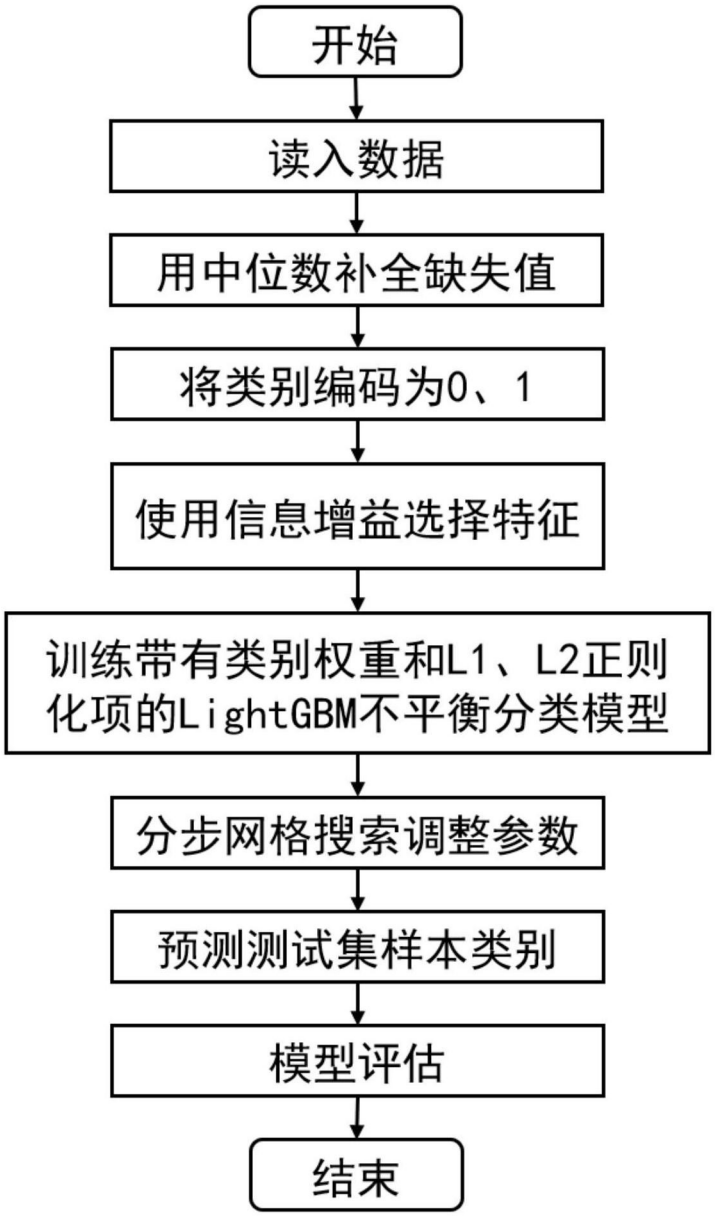

开始
读入数据
用中位数补全缺失值
将类别编码为0、1
使用信息增益选择特征
训练带有类别权重和L1、L2正则化项的LightGBM不平衡分类模型
分步网格搜索调整参数
预测测试集样本类别
模型评估
结束

图1

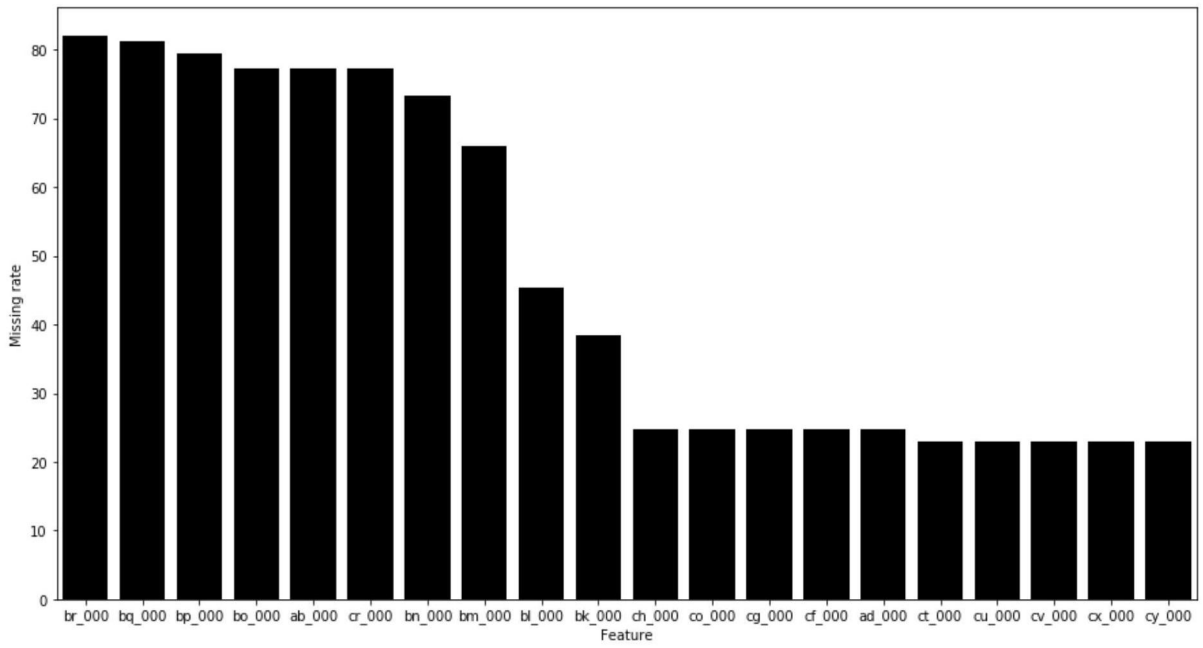

Missing rate
0
10
20
30
40
50
60
70
80
br_000 bq_000 bp_000 bo_000 ab_000 cr_000 bn_000 bm_000 bl_000 bk_000 ch_000 co_000 cg_000 cf_000 ad_000 ct_000 cu_000 cv_000 cx_000 cy_000
Feature

图2

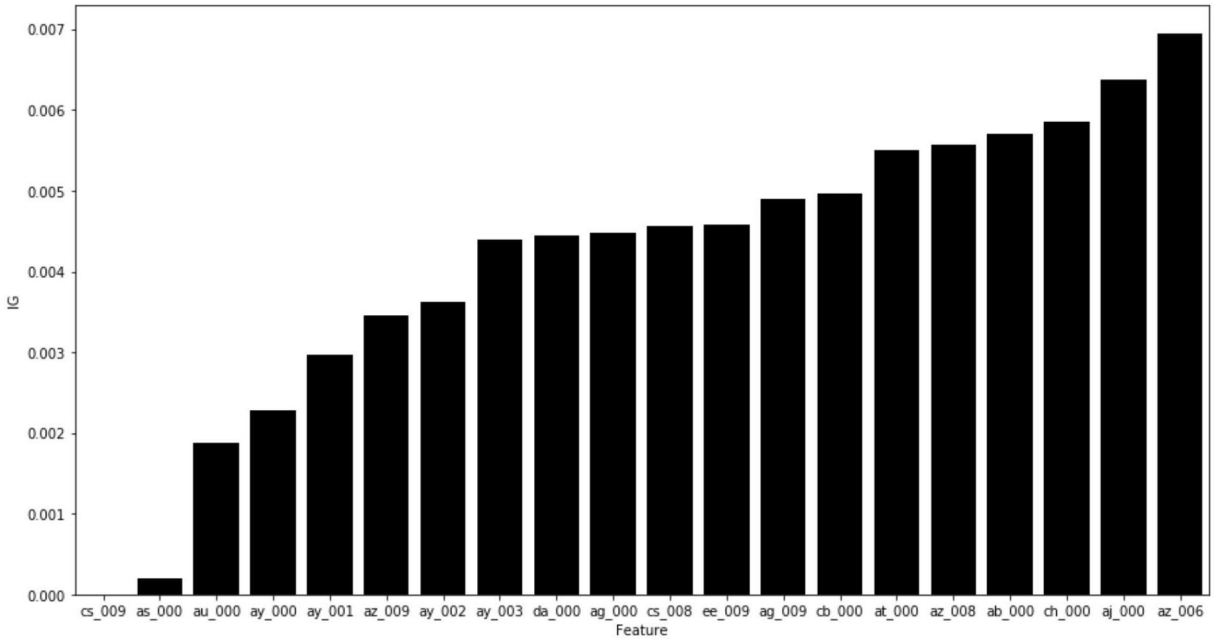

IG
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
cs_009 as_000 au_000 ay_000 ay_001 az_009 ay_002 ay_003 da_000 ag_000 cs_008 ee_009 ag_009 cb_000 at_000 az_008 ab_000 ch_000 aj_000 az_006
Feature

图3