(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116128340 A
(43)申请公布日 2023.05.16

(21)申请号 202211664674.2

(22)申请日 2022.12.23

(71)申请人 杭州电子科技大学
地址 310018 浙江省杭州市杭州经济技术开发区白杨街道2号大街1158号

(72)发明人 刘俊 徐浩浩 谷雨 陈华杰 彭冬亮

(74)专利代理机构 浙江永鼎律师事务所 33233
专利代理师 周希良

(51)Int.Cl.
G06Q 10/0639(2023.01)
G06Q 50/26(2012.01)
G06F 18/2431(2023.01)
G06F 18/214(2023.01)

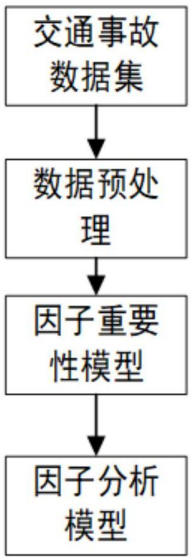
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种交通事故的影响因子分析方法

(57)摘要

本发明涉及一种交通事故的影响因子分析方法，包括：采集交通事故数据集，交通事故数据集包括交通事故的各种影响因子；对交通事故数据集进行数据预处理，得到目标交通事故数据集；利用随机森林对目标交通事故数据集的影响因子进行重要性评估，得到影响因子的重要性评分；按照重要性评分由小到大进行排序作为K2算法模型的输入顺序，将目标交通事故数据集输入K2算法模型进行训练，剔除与交通事故无关的影响因子，剩余的影响因子构成新的数据集；将新的数据集重新输入K2算法模型进行训练，得到影响因子分析模型，并利用影响因子分析模型进行影响因子分析。本发明解决了交通事故的影响因子分析准确率不高的缺陷，弥补了贝叶斯网络模型的不足。



CN 116128340 A

1.一种交通事故的影响因子分析方法,其特征在于,包括以下步骤:

S1、采集交通事故数据集,交通事故数据集包括交通事故的各种影响因子;

S2、对交通事故数据集进行数据预处理,得到目标交通事故数据集;

S3、利用随机森林对目标交通事故数据集的影响因子进行重要性评估,得到影响因子的重要性评分;

S4、按照重要性评分由小到大进行排序作为K2算法模型的输入顺序,将目标交通事故数据集输入K2算法模型进行训练,剔除与交通事故无关的影响因子,剩余的影响因子构成新的数据集;

S5、将新的数据集重新输入K2算法模型进行训练,得到影响因子分析模型,并利用影响因子分析模型进行影响因子分析。

2.根据权利要求1所述的影响因子分析方法,其特征在于,所述步骤S5中,利用影响因子分析模型进行影响因子分析,得到对交通事故直接影响的重要影响因子。

3.根据权利要求1所述的影响因子分析方法,其特征在于,还包括:

基于目标交通事故数据集的影响因子,利用贝叶斯估计法对其造成交通事故死亡的概率进行估计。

4.根据权利要求3所述的影响因子分析方法,其特征在于,还包括:

判断影响因子造成交通事故死亡的概率是否超出预设阈值;若是,则将相应影响因子造成交通事故死亡的概率设置为1。

5.根据权利要求1所述的影响因子分析方法,其特征在于,所述预设阈值为0.9～0.99。

6.根据权利要求1所述的影响因子分析方法,其特征在于,所述步骤S1中,设置数据采集表,表列为影响因子,包括交通事故周围的场景、时间、地点、环境、事故人数、伤亡程度;

根据数据采集表采集交通事故数据集。

7.根据权利要求6所述的影响因子分析方法,其特征在于,所述步骤S2,具体包括以下步骤:

S21、将交通事故数据集中的缺失数据删除;

S22、对交通事故数据集中的异常数据进行替换处理,得到新的交通事故数据集;

S23、对新的交通事故数据集进行离散化处理以及区间划分,得到目标交通事故数据集。

8.根据权利要求7所述的影响因子分析方法,其特征在于,所述步骤S22中,基于3σ准则的中值方法对异常数据q_i替换为处理过的数据l_i;其中,若$|vb|=|q_i-x|>3\sigma$,则为异常数据q_i;

$$q_i = l_i = \frac{z}{n}$$

其中,n代表同一影响因子对应的数据量,z代表同一影响因子对应的正常数据累加值,x为同一影响因子对应的所有数据的算术平均值,vb为异常数据q_i的剩余误差,σ为同一影响因子对应的所有数据的标准差。

9.根据权利要求1所述的影响因子分析方法,其特征在于,还包括:

将目标交通事故数据集的影响因子划分为可控因子和不可控因子,可控因子为人为可控制的,不可控因子为偶然因素。

一种交通事故的影响因子分析方法

技术领域

[0001] 本发明属于机器学习技术领域,涉及一种交通事故的影响因子分析方法。

背景技术

[0002] 现在,越来越多的选择汽车出行,而且道路里程的增加,路况的变好以及车辆驾驶路程的快速的增长,出行的交通频率和道路的交通流量也得到了明显的增大,但是交通事故频发,目前道路交通安全也成为了需要关注的安全,分析交通事故的影响因子是非常重要的。

[0003] 现有技术中的交通事故的影响因子分析方法主要存在以下三个问题:

[0004] 1、目前对于交通事故影响的因子的分析不够全面,考虑的因子太少。

[0005] 2、对于交通事故数据信息的采集没有统一的规格;

[0006] 3、目前对于交通事故的影响因子分析,传统模型有人工定义的主观性的误差,以及传统模型的精度太差,不能提供良好的影响因子分析。

发明内容

[0007] 基于现有技术存在的上述不足,本发明的目的是提供一种交通事故的影响因子分析方法,以解决目前没有良好的交通事故因子分析以及传统模型精度差的问题,提出了一种交通事故信息采集标准,解决目前交通事故数据采集不够统一的问题,构建了一种基于3σ准则的中值方法,除去异常数据,然后构建随机森林因子重要性模型自动在数据集中提取重要因子,然后提出了辨别因子重要度的方法,提高了传统贝叶斯模型的准确性,以及规避了人工参与的主观性,提出了基于因子重要性的影响因子分析模型来进行交通事故因子分析,并提高了交通事故影响因子分析的精度。

[0008] 为了实现上述发明目的,本发明采用如下技术方案:

[0009] 一种交通事故的影响因子分析方法,包括以下步骤:

[0010] S1、采集交通事故数据集,交通事故数据集包括交通事故的各种影响因子;

[0011] S2、对交通事故数据集进行数据预处理,得到目标交通事故数据集;

[0012] S3、利用随机森林对目标交通事故数据集的影响因子进行重要性评估,得到影响因子的重要性评分;

[0013] S4、按照重要性评分由小到大进行排序作为K2算法模型的输入顺序,将目标交通事故数据集输入K2算法模型进行训练,剔除与交通事故无关的影响因子,剩余的影响因子构成新的数据集;

[0014] S5、将新的数据集重新输入K2算法模型进行训练,得到影响因子分析模型,并利用影响因子分析模型进行影响因子分析。

[0015] 作为优选方案,所述步骤S5中,利用影响因子分析模型进行影响因子分析,得到对交通事故直接影响的重要影响因子。

[0016] 作为优选方案,影响因子分析方法,还包括:

[0017] 基于目标交通事故数据集的影响因子，利用贝叶斯估计法对其造成交通事故死亡的概率进行估计。

[0018] 作为优选方案，影响因子分析方法，还包括：

[0019] 判断影响因子造成交通事故死亡的概率是否超出预设阈值；若是，则将相应影响因子造成交通事故死亡的概率设置为1。

[0020] 作为优选方案，所述预设阈值为0.9～0.99。

[0021] 作为优选方案，所述步骤S1中，设置数据采集表，表列为影响因子，包括交通事故周围的场景、时间、地点、环境、事故人数、伤亡程度；

[0022] 根据数据采集表采集交通事故数据集。

[0023] 作为优选方案，所述步骤S2，具体包括以下步骤：

[0024] S21、将交通事故数据集中的缺失数据删除；

[0025] S22、对交通事故数据集中的异常数据进行替换处理，得到新的交通事故数据集；

[0026] S23、对新的交通事故数据集进行离散化处理以及区间划分，得到目标交通事故数据集。

[0027] 作为优选方案，所述步骤S22中，基于3σ准则的中值方法对异常数据q_i替换为处理过的数据l_i；其中，若$|vb|=|q_i-x|>3\sigma$，则为异常数据q_i；

[0028] $$q_i = l_i = \frac{z}{n}$$

[0029] 其中，n代表同一影响因子对应的数据量，z代表同一影响因子对应的正常数据累加值，x为同一影响因子对应的所有数据的算术平均值，vb为异常数据q_i的剩余误差，σ为同一影响因子对应的所有数据的标准差。

[0030] 作为优选方案，影响因子分析方法，还包括：

[0031] 将目标交通事故数据集的影响因子划分为可控因子和不可控因子，可控因子为人为可控制的，不可控因子为偶然因素。

[0032] 与现有技术相比较，本发明具有的有益效果为：

[0033] 本发明通过制定了交通事故信息采集标准，使得数据的采集更加全面，更加科学，以及提出了一种优化后的等间距法，解决了交通事故数据采集标准化和数据离散化适应性不够的问题，解决了全面分析交通事故的影响因子的问题；

[0034] 本发明使用随机森林的因子重要度结合因子分析模型对交通事故进行建模，能够更好的学习交通事故的发生规律以及数据特征，提高了系统的自适应能力；本发明通过因子重要度方法进行特征重要性比较，避免了人工设计重要性对场景泛化能力不足的问题；

[0035] 本发明构建了基于3σ准则的中值方法，提出了一种处理事故异常数据的数据预办法，可以发现异常数据并直接进行异常替换，有效的减少了数据异常对分析结果的影响，提高了模型的稳定性；

[0036] 本发明的影响因子分析模型，将因子重要性和因子关联分析两大优点，更好地对交通事故数据进行分析，新的模型对网络节点之间的相互作用以及依赖关系进行挖掘，更为精确的划分了交通事故因子之间的关系，使得出的因子网络结构更为科学合理；一定程度上解决了交通事故影响因子分析准确率不高，弥补了传统交通事故分析模型自身的不足的问题。

附图说明

[0037] 图1为本发明实施例的交通事故的影响因子分析方法的流程图；

[0038] 图2为本发明实施例的数据采集及数据预处理的示意图；

[0039] 图3为本发明实施例的影响因子重要性的分析示意图。

具体实施方式

[0040] 为了更清楚地说明本发明实施例，下面将对照附图说明本发明的具体实施方式。显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图，并获得其他的实施方式。

[0041] 如图1所示，本发明实施例的交通事故的影响因子分析方法，包括以下步骤：

[0042] S1、采集交通事故数据集，交通事故数据集包括交通事故的各种影响因子。

[0043] 具体地，设置数据采集表，表列为影响因子，包括交通事故周围的场景、时间、地点、环境、事故人数、伤亡程度等；其中包含各种环境因子、人为因子等数据，在事故发生后可以采用此数据采集表来搜集数据，为以后数据分析提供方便，也可以找相关部门直接获取原始的交通事故数据集，作为原始数据。

[0044] 因此，根据数据采集表采集交通事故数据集。

[0045] S2、对交通事故数据集进行数据预处理，得到目标交通事故数据集。

[0046] 如图2所示，上述步骤S2具体包括以下步骤：

[0047] S21、将交通事故数据集中的缺失数据删除；

[0048] S22、对交通事故数据集中的异常数据进行替换处理，得到新的交通事故数据集；

[0049] 具体地，基于3σ准则的中值方法对异常数据q_i替换为处理过的数据l_i；其中，若$|vb|=|q_i-x|>3\sigma$，则为异常数据q_i；

[0050] $$q_i = l_i = \frac{z}{n}$$

[0051] 其中，n代表同一影响因子对应的数据量，z代表同一影响因子对应的正常数据累加值，x为同一影响因子对应的所有数据的算术平均值，vb为异常数据q_i的剩余误差，σ为同一影响因子对应的所有数据的标准差。

[0052] 本发明实施例先对各个影响因子的数据点进行数值异常处理，先将异常点删除，比如在数据集中有数据显示没有采集的数据，以及采取的数据时异常的。采取的办法是先将原有异常点进行删除，再使用插入中值的方法将失去的点近似还原，这样可以把因素的异常数据进行了处理分析，构建成基于3σ准则的中值方法的模型。

[0053] S23、对新的交通事故数据集进行离散化处理以及区间划分，得到目标交通事故数据集。

[0054] 进行处理分析后的数据，即新的交通事故数据集，进行离散化处理，如果数据本身为离散数据，不再进行数据离散化，离散的同时应将数据进行区间划分，划分的时候应遵循平均原则，使数据的划分更加合理，按照优化后的等距法来进行区间划分，原有的等间距法只是平均的分为几个区间，将数据等间距划分，优化后的等间距法结合经验对于某些因子特征进行干预，更加科学，离散化完成后形成新的数据集，即目标交通事故数据集。其中，离

散化处理以及区间划分具体过程可参考现有技术，在此不赘述。

[0055] 另外，本发明还构建了交通事故信息分类标准，根据交通事故收集到的信息，可以将影响因子划分为可控因子和不可控因子，可控的为人为可以控制的因子，如速度，时间等，不可控的偶然因素，有突发疾病，车辆故障等。接下来，如图3所示，继续步骤S3。

[0056] S3、利用随机森林对目标交通事故数据集的影响因子进行重要性评估，得到影响因子的重要性评分。

[0057] 具体地，可利用随机森林的CART树模型，基于CART树模型的训练过程，利用平均不纯度减少法评估每个影响因子的重要性，得到影响因子的重要性评分。

[0058] 其中，影响因子的重要性评分的评估过程包括：

[0059] 首先，在目标交通事故数据集中随机抽取N个样本，即在目标交通事故数据集中随机找到一个样本，然后将样本再放回去，抽取总共N个样本后，停止采样，建立W个决策树，需要重复W次；在每个被抽取的样本集合上面，来创建W棵决策树；在构建好决策树后，在每个影响因子数据集中，随机抽取数据，作为训练集，进行训练；构建好决策树以后，然后从每个抽取样本的特征中随机抽取特征，然后来计算基尼指数，查看哪个因素对输出变量影响最大。具体地，根据数据集来构造函数，来计算每个特征X_j的Gini指数VIM_j，亦即第j个特征在RF所有决策树中节点分裂不纯度的平均改变量。第i棵树节点q的Gini指数的计算公式为：

[0060] $$GI_q = 1 - \sum_{c=1}^{|C|} p_{qc}{}^2$$

[0061] 其中，C表示有C类别，p_{qc}表示节点q中类别c所占的比例。如果，特征X_j在决策树中出现的节点为集合为M，则X_j在第i颗树的重要性为

[0062] $$VIM_j^{(i)} = \sum_{q \in M} VIM_{jq}^{(i)}$$

[0063] 设置有W个决策树，则特征X_j的Gini指数评分：

[0064] $$VIM_j = \sum_{i=1}^{W} VIM_j^{(i)}$$

[0065] 最后进行归一化的处理：

[0066] $$VIM_j = \frac{VIM_j}{\sum_{z=1}^{j} VIM_z}$$

[0067] 则可以得到对输出变量的重要性评分。

[0068] 上述重要性评分的评估过程具体可参考现有技术，在此不赘述。

[0069] 在计算得到每个影响因子的重要性评分后，从小到大建立评分顺序集合D，此集合为输入顺序集合，代表K2算法的输入顺序的重要程度。

[0070] S4、按照重要性评分由小到大进行排序作为K2算法模型的输入顺序，将目标交通事故数据集输入K2算法模型进行训练，剔除与交通事故无关的影响因子，剩余的影响因子构成新的数据集。

[0071] 具体地，将目标交通事故数据集输入K2算法模型进行训练，得到初步的影响因子

分析模型;由此模型可得到与交通事故无关的影响因子,剔除与交通事故无关的影响因子,剩余的影响因子构成新的数据集。

[0072] S5、将新的数据集重新输入K2算法模型进行训练,得到影响因子分析模型,并利用影响因子分析模型进行影响因子分析。

[0073] 根据构建的影响因子分析模型,可以得到交通事故的重要因子,即对交通事故有直接影响的因素;次要因素放到次要因子集合E里面。

[0074] 另外,本发明实施例还根据目标交通事故数据集的影响因子,使用贝叶斯估计法估计其造成交通事故死亡的概率;结合经验定义,传统的贝叶斯估计不够灵活,在面对特定场景时,适应性不够,结合经验优化,可以得到量化后的因果模型。

[0075] 具体地,判断影响因子造成交通事故死亡的概率是否超出预设阈值,预设阈值为0.9～0.99,具体可根据实际需求进行确定;若是,则将相应影响因子造成交通事故死亡的概率设置为1,以此来提升交通安全的可靠性。

[0076] 本发明实施例的交通事故的影响因子分析方法,具有如下优点:

[0077] (1)制定了交通事故信息采集标准,使得数据的采集更加全面,更加科学;

[0078] (2)提出了一种优化后的等间距法,解决了交通事故数据采集标准化和数据离散化适应性不够的问题,解决了全面分析交通事故的影响因子的问题;

[0079] (3)构建了基于3σ准则的中值方法,提出了一种处理事故异常数据的数据预办法,可以发现异常数据并直接进行异常替换,有效的减少了数据异常对分析结果的影响,提高了模型的稳定性。

[0080] (4)通过利用随机森林结合K2算法构建影响因子分析模型,对网络节点之间的相互作用以及依赖关系进行挖掘,更为精确的划分了交通事故与影响因子之间的关系,使最终的因子网络结构更为科学合理;一定程度上解决了交通事故影响因子分析准确率不高,弥补了贝叶斯网络模型自身的不足的问题。

[0081] 以上所述仅是对本发明的优选实施例及原理进行了详细说明,对本领域的普通技术人员而言,依据本发明提供的思想,在具体实施方式上会有改变之处,而这些改变也应视为本发明的保护范围。

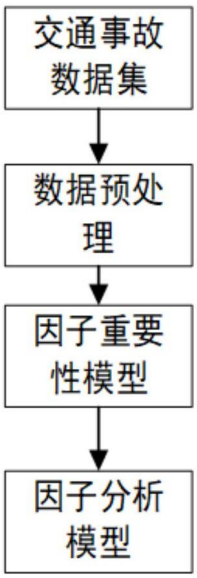

交通事故数据集
数据预处理
因子重要性模型
因子分析模型

图1

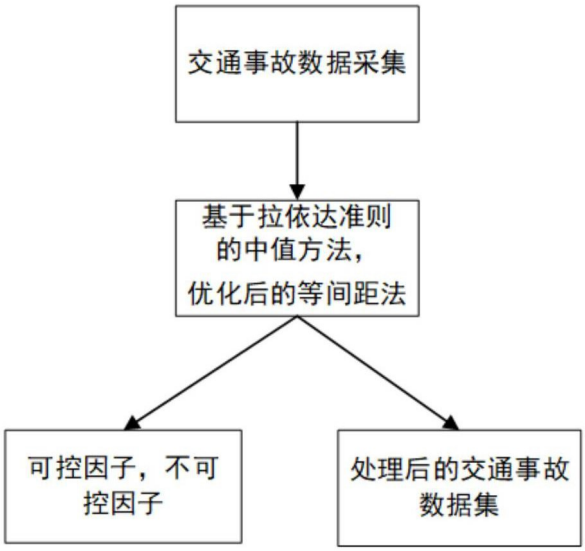

交通事故数据采集
基于拉依达准则的中值方法，优化后的等间距法
可控因子，不可控因子
处理后的交通事故数据集

图2

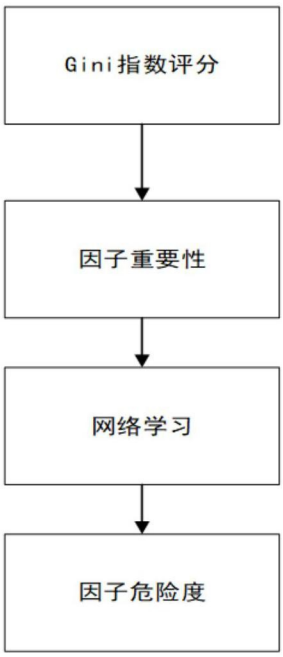


图3