基于决策树的轨道交通安全评估方法及其应用*

蔡国强[1**] 贾利民[1] 吕晓艳[2] 刘春煌[2]

1. 北京交通大学轨道交通控制与安全国家重点实验室, 北京 100044
2. 铁道科学研究院电子计算技术研究所, 北京 100081

摘要 提出了一种定量的基于决策树的轨道交通安全评估方法. 该方法针对风险源的海量数据集和庞大的属性集, 采用面向应用的属性构造、规范化与面向属性的归约进行数据集的预处理, 为突破内存限制, 采用改进的决策树分类方法 DT _ SA 对轨道交通风险因素进行分析, 最终提取的分类规则包含描述类分布的定量信息. 同时规则前件中属性的顺序又体现了属性对主类的影响程度. 这些规则揭示了轨道交通风险源的规律, 应用验证了其有效性.

关键词 轨道交通 安全评估 决策树

安全是轨道交通永恒的主题. 长期以来, 世界各国为确保轨道交通运输安全已做了大量的工作. 但是, 全世界列车脱轨、撞车、火灾等等事故仍屡屡发生, 并造成了严重的后果. 究其原因, 除了轨道交通的快速发展和少数自然灾害不可抗力等原因外, 绝大部分都是由于车辆状态、营运控制、人员态度与技能等人为因素所造成的. 我国轨道交通依照“提速-改革-安全-再提速”的良性循环模式发展的同时安全问题也日益得到人们的重视. 为此, 国内外的安全领域专家开展了大量针对安全评估方法的研究工作[1—3], 并将一些先进的安全管理与评估方法应用到保障轨道交通运营安全之中.

结合对决策树[4—8]的研究, 本文分析了轨道交通安全评估的一般步骤, 并提出一种基于决策树的轨道交通定量安全评估方法 DT _ SA (decision tree _ safety analysiss), DT _ SA 是一种精确性较高的定量安全评价方法. 能够给出安全评价规则, 用于指导评价系统的危险性, 判定其是否达到预期的安全要求, 提供对系统安全性评定及制定安全措施的依据. 实际应用满足了货车提速可靠性试验的安全保障评估要求.

1 轨道交通安全评估

轨道交通安全评估是以实现运输安全为目的, 利用系统工程的理论和方法, 对风险源, 即安全隐患进行识别、分析和评价, 研究重大安全隐患的发生机理、判别标准, 制定有效的操作方案、操作程序、预防措施及应急补救措施, 对未来短、中长期的安全状况进行推断, 为安全指标制定、事故分析和安全标准规程的改善提供依据.

轨道交通安全评估的研究包括风险分析、安全评估和风险管理 3 个部分. 风险分析对风险源进行量化分析和研究; 安全评估指对风险本身的评估; 风险管理指具体风险系统监控和管理.

1.1 风险分析

风险分析的目标是研究不利后果是怎样产生的, 为什么会产生. 实际上是指对风险源进行识别分类、建模的过程. 通过掌握风险源的状态, 进行风险管理, 减小或控制风险. 风险分析包括两个环节:

2007-03-25 收稿, 2007-05-12 收修改稿

* 国家自然科学基金资助项目(批准号 60033202o, 60674001)

** E-mail: caiguoqiang@jtys.bjtu.edu.cn

(1) 风险源分析

风险源分析的主要任务是研究风险源的组成要素-风险源因子发生的概率或重现期. 任何风险源因子都需要 3 个参数才能加以完整的刻画, 即时、空、强[10].

时: 风险源因子出现或发生作用的时间(时间点或时间段).

空: 风险源因子所在的地理位置(可能是一个运行区间).

强: 风险源因子强度, 如车辆晃动级别, 车轮踏面擦伤级别.

为了从量化层次上对风险源因子进行分析, 首先要定义风险源因子的测度空间. 例如: 车轮擦伤级别的集合 {1 级、2 级、3 级} 就是车辆设备老化的一种测度空间.

风险的不确定性可以分为随机不确定性和模糊不确定性两种, 主要是由风险源因子决定的. 根据风险源的特点可以将其分为: 现实风险; 概率风险和预测风险 3 类.

(2) 事故损失评估

事故通常会使轨道交通的正常运行中断, 造成人员伤亡或财产损失等不良后果.

事故损失评估是对风险范围内一定时段内可能发生的一系列事故造成的损失进行评估.

(3) 风险分析具体步骤

风险分析的目标是描述掌握风险源状态, 以便进行风险管理, 减小或控制风险. 其具体步骤如图 1 所示:

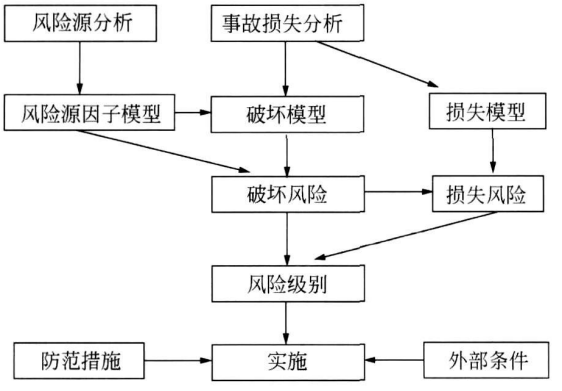

图 1　风险分析步骤

1.2　安全评估

安全评估的目的认识和描述风险源的特性, 根据各种风险模型来评估风险的影响途径及趋势, 制定系统框架, 以便避免和控制风险. 涉及的主要模型包括风险源因子模型、破坏模型和损失模型:

(1) 风险源因子模型

风险源因子模型可以用概率关系式: Prob(T, G, I)表示, 即在 T 时或时间段内, 在 G 区域上强度为 I 的风险源发生的概率. 但是, 许多风险源因子含有不确定性成分, 基于模糊集合理论的不确定因素建模方法是研究的热点[10—12].

(2) 破坏模型

破坏模型用来描述风险源因子强度和事故破坏程度之间的函数关系. 实际中这种函数关系很难找到, 因此, 破坏模型主要采用事故模拟, 根据历史数据统计的方法来寻找近似的函数.

(3) 损失模型

损失模型包括经济损失模型和人员损失模型, 主要基于统计的方法获得.

1.3　风险管理

风险管理是在安全评估的基础上, 制定风险源的监控对策和管理机制, 以消除或抑制风险源, 规避风险.

风险管理包括: 设计管理方案; 选择和实施管理系统; 监控和核查. 从监控风险主客观因素出发, 包括监控关键岗位运营作业操作行为; 监控车辆运行状态、车辆装载状态, 危险品远程运输过程; 监控基础设施状态以及轨道交通运输与其他运输方式安全交互等方面构成一套综合体系, 实现轨道交通安全相关过程的 "超前预警, 控制风险, 最小化损失", 促进安全保障系统完善实践.

2　基于决策树的轨道交通安全评估方法 DT _ SA

目前, 国内外安全评估方法的设计结构受行业特定的对象限制, 还没有形成满足普遍适用的评估方法. DT _ SA 方法是一种定量的安全评估方法, 该方法的步骤如下:

(1) 数据预处理;

(2) 构造评估决策树;

(3) 提取评估规则;

(4) 利用规则评估对象安全.

DT _ SA 的主要特点是:

(1) 采用基于文件分割的改进决策树方法, 以突破对数据集数据量的内存限制, 适应对大数据集的处理;

(2) 基于数据集的分割, 使算法具有较好的并行性;

(3) 叶结点包含类分布信息, 实现了定性与定量的结合.

2.1 数据预处理

数据预处理技术可以改进数据的质量, 提高分析的精度和性能. 针对不同的应用需求应用不同的预处理策略. 常用的预处理方法主要包括: 数据清理、数据融合、数据变换和数据归约. 数据清理去掉数据中的噪声, 纠正不一致. 数据融合将多种数据源合并成一个一致的数据存储, 如数据仓库或数据立方体. 数据变换操作是对数据的规格化和聚集. 数据归约通过聚集、删除冗余特性或聚类等方法来压缩数据. 这些数据处理方法在预测模型建立之前使用, 可以大大提高模型的质量, 降低实际建模所花费的时间.

(1) 数据变换

数据变换是指将数据转换成适于高效处理的形式, 主要方法包括平滑、聚集、数据概化、规范化和属性构造. 其中数据概化是指使用概念分层, 用高层次概念替换低层次 "原始" 数据. 给定的属性进行构造和添加新的属性以帮助提高精度和对高维数据结构的理解. 利用此方法可以帮助平缓因使用决策树算法分类而导致的分裂问题.

(2) 数据规约

数据归约可以用来得到数据集的归约表示, 它相对于原数据集小得多, 但仍接近保持原数据的完整性. 这样, 对归约后的数据集进行处理将更有效, 并产生相同或近似相同的分析结果. 数据归约的策略包括数据立方体、聚集、维归约、数据压缩、数值压缩和离散化.

数据离散化主要是对连续属性进行的, 可以用来减少给定连续属性值的个数. 这种数据预处理对于使用基于判定树的分类挖掘方法有很多优点. 离散化主要以预测分析应用需求和决策者的需求来进行离散化.

2.2 构造评估决策树

构造评估决策树(construct analysis decision tree)的形式化描述如下文所述:

```
Construct_Analysis_Decision_Tree (S)
Input: safety data sets S
       Number of main class ci,
       Number of attributes ai in S,
       Number of each attribute distinct values k,
Output: safety analysis decision tree T
Method:
Initialize (T, S);  /* S is the data set of current node */
N= LeafNode(S);  /*a leaf node is generated */
For i= 1 to ci do
    If ‖S‖= ‖Ci‖  /*if all the data in S are labeled with
                    Ci */
    {  Label(N)= i;  /* the label of N is set to i */
       Mark(N)= leaf;  /* a leaf node is marked */
       Return T;
    }
A0= UnselectedAttrOnly(A);  /*A0 is the instance space
described only by unselected attributes */
If CannotDistinguish(S, A0);  /*if attributes cannot be
further selected and distinguished */
{
    Label(N)= the label decided by major voting;
    Return T;
}
ε= SelectSplit(S, A0);  /* a split attribute is selected ac-
cording to information gain ratio */
MarkselectedAttr[ε] = true ;  /* if a attribute in instance
space is selected as split attribute, marked it */
Split (ε, S) into S1, . . . , Sd;  /* several subsets are genera-
ted */
For k= 1 to d do
{   If ‖Sk‖= 0 {Child(N, k)= nil; continue;}
    Label (N)= ε;  /* the label of N is set toε */
    Mark (N)= not leaf;
    Child (N, k)= DP (Sk);  /* recursively process child
    of current node */
}
```

算法中，在构造评估决策树时，初始的训练数据集对应树的根结点．在建树过程中，每次从经过预处理的训练集中选择出相对于主类区分度最好的属性，以其为决策属性，根据数据在该属性下的取值对训练数据集投影分割，形成子训练数据集且对应不同的树结点；递归地对每个子训练数据集重复进行，最终，可以获得一棵多叉决策树．由于安全监测数据集的属性繁多，所得的决策树庞大，修剪决策树显得十分重要．针对实际的应用决策需求，不可将决策树进行过分的修剪以避免有用信息的丢失，特别是所含的样本数低于设定阈值的结点在剪枝前需要详细分析，因为这些样本中包含的往往是异常信息，而这些异常信息表达的是非平凡的知识，更具有指导意义．

上述方法构造的决策树与其他决策树构造方法的不同之处在于每个叶结点保留了关于各个类的分布信息，沿着各分支向上攀升，可以得到各个相应内结点的主类分布信息，这些包含了传统的统计汇总的信息，使得从决策树中提取的规则不仅含有定性信息，而且含有定量的信息．考虑到内部结点的类分布访问不是十分频繁，因此只在叶结点存储这些主类分布信息，这样只在需要时才通过简单的计算由叶结点推出其主类的分布，从而可以节省空间．

规则的提取原则和传统的规则提取一样，规则的前件是根结点到叶结点沿分支按顺序组成．各属性在决策树中的位置依次体现了其区分主类的重要性，越靠近根结点，此属性对于区分主类越有相对的重要性．不同之处在于提取的规则后件构成不仅包含分类的定性信息，而且包含主类分布信息的定量信息，实现了统计与预测的集成．下面给出了一个关于定量规则的描述：

左侧温＝75℃以上∧右侧温 50—75℃∧温度 5—25℃∧速度＝100 km/h 以上⇒左侧温升＝70—100℃．　（0.75%）

3　DT _ SA 的应用

3.1　评价样本

为满足提速货车的需要，铁道部进行了提速货车 120km/h 的可靠性试验．为保障试验的安全进行，研制了货车安全综合监控系统，对车辆走行部进行监控．经过 10 个月的运行，积累了大量的货车关键部件监测数据．运用 DT _ SA 方法处理了积累的红外轴温历史数据共 6605060 条，从中得出了有益的结论．

（1）试验车辆数据

试验车辆的信息如表 1：

表 1　试验车辆信息

车种车型	车种编号	空重别编码	空重别
P64GK	0	0	重
P65GK	0	1	半
P66GK	0	2	空
NX17KB	1	0	重
NX18KB	1	1	半
NX19KB	1	2	空
G70K	2	0	重
G70K	2	1	半
G70K	2	0	重
C64K	3	0	重
C64K	3	0	重
C64K	3	2	空
G70H	4	0	重
G70H	4	0	重
G70H	4	2	空
P64GH	5	0	重
P64GH	5	1	半
P64GH	5	2	空
NX17BH	6	0	重
NX17BH	6	1	半
NX17BH	6	2	空
C64H	7	0	重
C64H	7	1	半
C64H	7	2	空
G70H	8	2	空
C80	9	0	重
C80	9	0	重
C80	9	2	空

(2) 数据属性

分析的数据属性如表 2 所示:

表 2 数据属性

标识	属性	描述	类型
1	SVEHICLENO	车号	varchar
2	sMonNodeCode	监测站编号	varchar
3	DPassTime	通过时间	datetime
4	FAveSpeed	速度	numeric
5	fTemperature	环境温度	numeric
6	sVehicleType	空重别	varchar
7	fLTemp	左侧温度	numeric
8	fLTempRise	左侧温升	numeric
9	fRTemp	右侧温度	numeric
10	fRTempRise	右侧温升	numeric

以分析左侧温升为例, 对选定数据集进行相关预处理后, 得到的训练集如表 3 所示:

表 3 左侧轴温分析训练数据集

类属	属性	属性值域
Main class	fLTempRise	0. 0—55℃
		1. 55—70℃
		2. 70—100℃
Attributes	FAveSpeed	0. < 80km/h
		1. 80—100km/h
		2. > 100km/h
	fTemperature	0. < 5℃ (冬天)
		1. 5—25℃ (春天)
		2. > 25℃ (夏天)
	fRTemp	0. < 0℃
		1. 0—20℃
		2. 20—30℃
		3. 30—40℃
		4. 40—50℃
		5. 50—75℃
		6. 75℃以上

3.2 轴温数据分析

采用 DT _ SA 方法进行数据分析的目的是针对影响轴承温升的各种因素进行分析, 得出评估规则, 制定合适的轴温报警域值, 评估货车的安全状态, 防止事故发生. 分析以左侧轴温为主属性进行, 建立左侧轴温与左侧温度、右侧温度、环境温度、右侧温升、速度之间的预测模型, 得出了较好的指导性结论.

实验分析的训练集如下:

(1) 目标类: 左侧温升

(2) 分类属性: 左侧温度、右侧温度、环境温度、右侧温升、速度

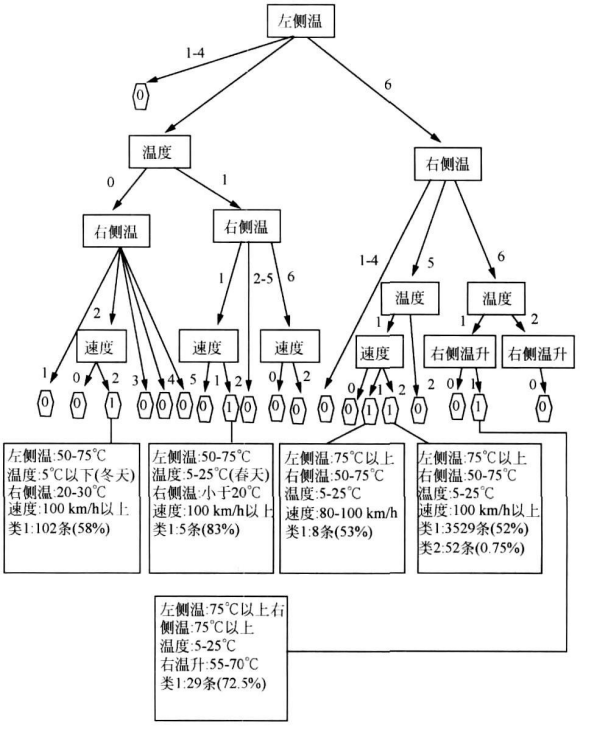

图 2 DT _ SA 构造的决策树

从图 2 中, 提取的类 1 和类 2 的 6 条定量规则如下:

(1) 左侧温=50—75℃∧温度=5℃以下(冬天)∧右侧温=20—30℃∧速度=100km/h 以上
⇒左侧温升=55—70℃ (58%)

(2) 左侧温=50—75℃∧温度=5—25℃(春天))∧右侧温<20℃∧速度=100 km/h 以上
⇒左侧温升=55—70℃ (83%)

(3) 左侧温=75℃以上∧右侧温 50—75℃∧温度 5—25℃∧速度=80—100 km/h
⇒左侧温升=55—70℃ (53%)

(4) 左侧温=75℃以上∧右侧温 50—75℃∧温度 5—25℃∧速度=100 km/h 以上
⇒左侧温升=55—70℃ (52%)

(5) 左侧温=75℃以上∧右侧温 50—75℃∧温度 5—25℃∧速度=100 km/h 以上
⇒左侧温升=70—100℃ (0.75%)

(6) 左侧温＝75 ℃以上∧右侧温 75 ℃以上∧温度 5—25 ℃∧右温升＝55—70
⇒左侧温升＝55—70 ℃　(72.5%)

3.3 评估规则

根据上面提取的规则，可以得到以下结论：

(1) 类 1 都是出现在高速时，说明速度显著影响轴承的温升. 速度为 80—100 km/h 时出现温升 75 ℃以上的比例很高，在这个速度区间内，轴承工作温度大都在温升 70—100 ℃，必须加强监控；

(2) 疲劳试验计划可以优化调整，在冬天温度低的情况下，温升较大概率在 70 ℃以下，可以适当增加运行时间；

(3) 根据规则 3，4，5 左右轴承温度温差应小于 20—25 ℃. 如果在高速时左右轴承温差介于 20—25 ℃之间，则温度高的轴承极可能是问题轴承，必须立即拦停检修，防止事故发生. 试验中的安全监控处理方法完全符合这条规则；

(4) 目前的轴温评判标准大都考虑同侧轴承的温升，依据量比和列比进行评判. 根据规则 3，4，5 的分析结果可知，同轴但不同侧的温升可以相互对比，作为评判的依据. 如果两侧温度差别大于 20—25 ℃，则可以简单有效地确定高温轴承存在问题，该实验结果可以修订目前的轴温评判标准.

4 结论

为了确保我国轨道交通的快速稳定发展，以及设计建造、营运管理等方面的安全工作与国际接轨，深入开展安全评估的研究至关重要. 本文给出的安全评估步骤和定量评估方法，建立了评估模型，并进行了实践尝试，评估结果为操作程序提供指导并为制定、修改相关安全标准与规定提供参考依据. 该评估方法可进一步应用于其他轨道交通安全保障系统中.

参 考 文 献

1 贾利民, 李　平. 铁路智能运输系统——体系框架预标准体系. 北京：中国铁道出版社. 2004, 8
2 方泉根, 王　津. A. Datubo. 综合安全评估(FSA) 及其在船舶安全中的应用. 中国航海. 2004, 1：1—5
3 杨建伟, 蔡国强. 车辆运行状态测试设备间相对误差率分析新方法, 计量学报, 2006, 27(1)：91—96
4 Quinlan JR. Induction of decision trees. In Machine Learning, 1986, 1：81—106
5 Wirth J, Catlett J. Experiments on the costs and benefits of windowing in ID3. In：5th Int'l Conference on Machine Learning, 1998
6 Quinlan JR. C4.5：Program for Machine Learning. San Matea CA：Morgan Kaufmann, 1993
7 Manish Mehta, Rakesh Agrawal, Jorma Rissanen. SLIQ：A fast scalable classifier for data mining. In：Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT'96), Avignon, France, March 1996
8 Shafer J, Agrawal R, Methta M. SPRINT：A scalable parallel classifier for data mining. Very Large Data Bases(VLDB'96), Valencia, Spain. 1996, 168—182
9 Cai GQ, Jia LM. Application research on track safety assessment based on artificial neural network, Proceedings of The 2007 International Conference on Information and Knowledge Engineering(IKE'07). 2007：25—28, Las Vegas, USA
10 黄崇复. 自然灾害风险分析的基本原理-自然灾害风险评价理论与实践. 北京：科学出版社, 2005
11 蔡国强, 贾利民, 刘春煌. 基于模糊穴的铁路风险分析及其在沪宁线中的应用. 中南大学学报(自然科学版), 2005, 36(1)：610—614
12 肖贵平. 铁路行车安全评价研究. 中国安全科学学报. 1995.8, (4)：32—36