列车仿真技术中基于属性矩阵图的故障分析决策树算法

王　超

（中国铁道科学研究院通信信号研究所，100081，北京//助理研究员）

摘　要　根据数据挖掘技术分析列车运行大数据的特点，提出了基于属性矩阵图的决策树算法。结合某列车仿真数据，详细阐述了计算属性度量、构建属性矩阵图模型及构造决策树的具体过程。由该决策树算法的故障分析结果可见，基于属性矩阵图决策树算法能准确地对故障问题进行分类归纳，为故障预测提供可靠依据。

关键词　属性矩阵图；决策树算法；列车仿真；故障分析

中图分类号　N945.25；U391.99：U27

DOI：10.16037/j.1007-869x.2017.12.025

Application of Decision Tree Optimization Algorithm in Train Simulation Technology

WANG Chao

Abstract　The data mining technology is used to analyze the large data generated during train operation，the decision tree algorithm is proposed based on attribute matrix graph. Combined with the simulation date of a train，the computing attribute matrix and the structure design of the decision tree optimization algorithm are elaborated. According to fault analysis result of the decision tree algorithm，this algorithm could classify the faults accurately and provide reliable basis for the prediction of metro faults.

Key words　attribute matrix graph；decision tree algorithm；train simulation；fault analysis

Author′s address　Signal & Communication Research Institute，China Academy of Railway Sciences，100081，Beijing，China

列车运行时，其车载设备每时每刻都要产生大量的数据。传统数据处理方法是先由车载设备存储日记，再由人工对文本格式的日志进行下载，这即便耗费了大量的时间和精力，也只是分析了部分数据。因此有必要引入数据挖掘技术，通过决策树模型在线分析处理列车运行数据，发现其中的关联规则。这不仅能尽早发现列车存在的故障隐患，提高列车的运行效率，也能节省人工核对数据的成本，具有非常高的现实意义。为此，提出了基于列车仿真技术的数据挖掘系统方案。

1　数据挖掘系统

列车仿真系统的工作原理如图 1 所示。

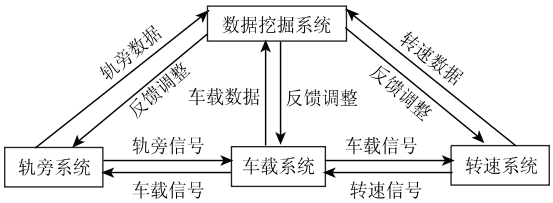

图 1　列车仿真系统原理图

由图 1 可见，数据挖掘系统是列车仿真系统的核心部分之一，主要负责处理来自于其他子系统的数据信息。通过反馈调整信息，并传输给各个子系统模块，可有效地减少故障发生的概率，提高列车运行的效率。数据挖掘系统解决了传统应用中对列车日志分析存在的重复耗时低效的问题，是一种自适应的智能学习系统。数据挖掘系统采用的决策树算法决定了其处理能力的强弱。

2　决策树算法

2.1　ID3 算法

ID3 算法是经典的决策树算法，其核心是采用信息熵和信息增益的方法来划分最佳决策树分裂点。该算法也存在着问题：首先，ID3 算法需要通过重复遍历数据集来计算每个属性的信息增益，故当数据集很大时，计算耗时会呈几何级数量增长；其次，ID3 算法不能对决策树进行动态更新，处理实时数据时易造成预测信息增益的偏差；最后，ID3 算法只能用来处理属性元素为离散变量的问题。

2.2　基于属性矩阵图的决策树算法

本文提出一种基于属性矩阵图决策树算法，改进了 ID3 算法存在的问题。利用属性矩阵图决策树算法能找出故障模式的规律，可发现列车运行过程中存在的隐患，能有效提高列车运行的安全性；通过动态构造决策树算法，可实时处理列车运行的数据，

能对可能遇到的故障模式进行预判。

2.2.1 构建属性矩阵图

对属性节点的划分决定了数据集的分裂方式。故基于属性矩阵图的决策树算法只有实现对属性节点合理划分，才能对数据规则进行挖掘和预测。

划分属性节点时，以属性度量来表示给予每个属性的评价，只有获得最好属性度量的属性才可作为分裂属性。根据信息论，期望信息越小，信息增益就越大，相应的分裂属性对确定整个系统划分的作用就越大，所以采用熵值和信息增益来进行属性度量。此处的熵值为整个数据集中属性的不确定性。令 x 表示对数据集划分的属性不确定性集合，则 x 的熵值定义为 $E(x)=-\sum_{n=1}^{m}(p_n \log_2 p_n)$：

对于第 i 类属性 x_i 有：

$$E(x|x)=-\sum_{i=1}^{m}\frac{x_i}{x}f(x),i\in(1,2,\cdots,m)$$

式中：

$E(x|x)$ ——属性对整个系统的条件熵值；

$f(x)$ ——数据集中的平均信息量；

$\frac{x_i}{x},i\in(1,2,\cdots,m)$ ——属性 x_i 在数据集 x 中所占的概率。

则属性 x_i 对整个数据集的信息增益为：

$g(x)=E(x)-E(x|x),i\in(1,2,\cdots,m)$

现截取部分列车仿真平台处理的车载数据，如表 1 所示。决策树算法的核心问题就是分析故障数据，发现其中的规律，并对故障进行分析和预测。

由表 1，经计算可得 $E(x)=0.880$。表 1 中列车保护速度及列车实际速度为连续型数据，其他均为离散型数据。

由表 1，列车 ID（标识）属性 X_i 按 0x01、0x02、0x03 分别取 x_1、x_2、x_3，则相应的条件熵值为：$E(x_1)=0.845$，$E(x_2)=0.811$，$E(x_3)=0.971$；故有 $E(x_1|x)=0.860$。列车 ID 属性对数据集的信息增益 g（列车 ID）$=E(x)-E(x|x)=0.020$。

同理，可分别算出其他离散型数据属性（控制、驾驶、信标 ID、SRP 及 BTM）的信息增益分别为：g（控制）$=0.000\ 5$；g（驾驶）$=0.030\ 0$；g（信标 ID）$=0.019\ 5$；g（SRP）$=0.016\ 0$；g（BTM）$=0.001\ 8$。

列车运行速度为连续型数据。本文采用一种基于速度窗口的方法来计算连续型数据属性的信息增益。根据相关行业规范，根据不同的列车运行等级和运行模式，可将列车的运行速度划分成不同等级。列车在不同运行速度等级下发生的故障往往具有类型一致性。根据这种特性，把列车的保护速度值和

表 1 部分列车车载设备仿真数据

列车ID	控制模式	驾驶模式	列车保护速度/（km/h）	列车实际速度/（km/h）	信标 ID	SRP	BTM	状态
0x01	ITC	RM	25.0	5.3	0x1a	激活	激活	正常
0x01	ITC	RM	25.0	6.2	0x1a	未激活	激活	正常
0x01	ITC	PM	25.0	7.1	0x1a	未激活	激活	故障
0x01	ITC	RM	25.0	8.3	0x1a	激活	未激活	正常
0x01	ITC	RM	25.0	6.0	0x1a	激活	激活	正常
0x01	ITC	RM	25.0	7.2	0x1a	激活	未激活	正常
0x01	ITC	AM	25.0	8.5	0x1b	激活	未激活	故障
0x01	CTC	PM	80.0	30.5	0x1c	激活	激活	正常
0x01	ITC	PM	60.0	35.4	0x1c	激活	激活	正常
0x01	ITC	PM	25.0	16.3	0x1c	激活	激活	正常
0x01	CTC	AM	60.0	60.6	0x1c	激活	激活	故障
0x02	ITC	RM	25.0	26.5	0x1b	激活	激活	故障
0x02	CTC	AM	80.0	60.2	0x1a	激活	激活	正常
0x02	CTC	PM	80.0	50.1	0x1a	激活	激活	正常
0x02	CTC	AM	60.0	50.8	0x1b	激活	激活	正常
0x03	ITC	RM	25.0	10.6	0x1b	激活	未激活	正常
0x03	ITC	RM	25.0	10.7	0x1b	激活	激活	正常
0x03	ITC	RM	25.0	26.9	0x1a	激活	激活	故障
0x03	CTC	PM	80.0	30.9	0x1b	激活	激活	正常
0x03	CTC	PM	80.0	70.2	0x1c	激活	激活	故障

注：ITC——点式控制等级；CTC——连续式控制等级，RM——受控人工驾驶模式；AM——自动驾驶模式；PM——防护人工驾驶模式；BTM——应答器传输单元

实际速度值分为 3 个速度窗口（速度单位为 km/h）。在相应的速度窗口内，速度具有相同的属性类别。由此计算可得 g（保护速度）=0.016 5 km/h，g（实际速度）=0.040 0 km/h。

根据上述计算结果可见，列车实际速度的信息增益最大，因此，选取实际速度作为数据集的分裂点。为便于计算属性的信息增益，需建立属性矩阵图模型以快速确定属性类别的状态和数量。根据表 1 数据，以列车实际速度作为数据集分裂点，构建属性矩阵图模型如图 2 所示。

根据属性矩阵的对应关系，可继续计算下层节点分裂属性的信息增益。现以列车实际速度在 [0，25] 区间的数据集为例进行计算。根据动态信息图，可以快速计算得出列车实际速度 ∈ [0，25] 数据集的熵值 E（实际速度 [0，25]）=0.720。

通过快速定位属性矩阵图中元素的统计值，可计算出在 [0，25] 内各属性的条件熵值为 E（0x1a）=0.65，E（0x1b）=0.91，E（信标 ID）=0.663，信标 ID 的信息增益 g（信标 ID）= E（实际速度 ∈ [0，25]）−E（信标 ID）=0.057；

以此类推，其余属性的信息增益分别为：g（驾驶等级）=0.020，g（信标 ID）=0.057，g（列车 ID）=0.071，g（控制等级）=0，所以列车实际速度 [0，25] 的分裂属性为列车 ID。

	0x1a	0x1b	0x1c	ITC	CTC	RM	PM	AM
[0 25]	正常：5 故障：1	正常：2 故障：1	正常：1 故障：0	正常：8 故障：2	正常：0 故障：0	正常：5 故障：2	正常：1 故障：1	正常：0 故障：1
(25 60]	正常：1 故障：1	正常：2 故障：1	正常：2 故障：0	正常：1 故障：2	正常：4 故障：1	正常：0 故障：2	正常：4 故障：0	正常：1 故障：1
(60 80]	正常：1 故障：0	正常：0 故障：0	正常：0 故障：2	正常：0 故障：0	正常：1 故障：2	正常：0 故障：0	正常：0 故障：1	正常：1 故障：1

图 2　根据案例数据建立的属性矩阵图

2.2.2　构造决策树，进行故障分析

根据此方法依次计算信息增益判断新的分裂点，构造决策树如图 3 所示。

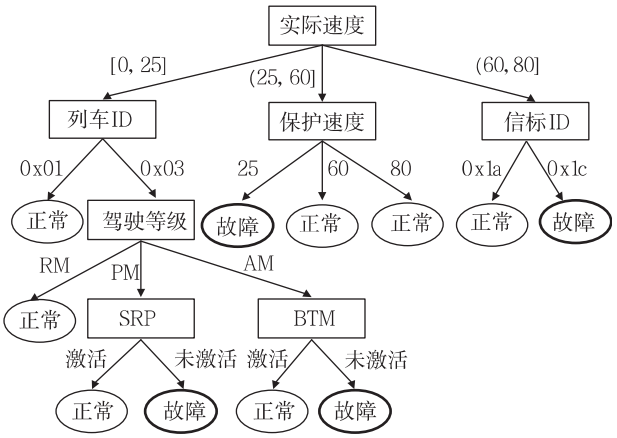

图 3　根据仿真数据构建的决策树

根据数据集构造的决策树，可以得到 4 类故障分析结果：

（1）列车实际速度在 [0，25] 的范围内，驾驶等级为 PM 模式，SRP 未激活。

（2）列车实际速度在 [0，25] 的范围内，驾驶等级为 AM 模式。

（3）列车实际速度在 (25，60] 的范围内，列车保护速度为 25。

（4）列车实际速度在 (60，80] 的范围内，信标 ID 为 0x1c。

可见，基于属性矩阵图决策树算法能准确地对故障问题进行分类归纳，为故障预测提供可靠依据。

3　结语

数据挖掘技术现已广泛应用在多个领域。城市轨道交通行业也在探索数据挖掘技术的应用方向。本文首次以数据挖掘技术为基础，针对城市轨道交通列车运行中的大数据问题，提出了属性矩阵图决策树算法，能准确地对故障问题进行分类归纳，为故障预测提供可靠依据。

参考文献

[1] 王威.基于决策树的数据挖掘算法优化研究 [J].现代计算机，2012，(19)：11.

[2] 王大玲，于戈，王国仁.基于概念层次树的数据挖掘算法的研究与实现 [J].计算机科学，2001，28 (6)：88.

[3] 胡笑蕾，胡华平，宋世杰.数据挖掘算法在入侵检测系统中的应用 [J].计算机应用研究，2004，21 (7)：88.

[4] 李良俊，张斌，杨明.一种基于模糊神经网络的数据挖掘算法 [J].计算机工程，2007，33 (12)：63.

[5] 孙亚，钱洪波，叶亮.数据挖掘算法在交通状态量化及识别的应用 [J].计算机应用，2008，28 (3)：738.

（收稿日期：2016−05−25）