DOI:10.16515/j.cnki.32-1722/n.2021.02.002

改进的 Apriori 算法在集中告警系统中的应用研究

李文锋[1],闫 涛[2]

(1.金陵科技学院软件工程学院,江苏 南京 211169;2.中国铁路通信信号上海工程局集团有限公司,上海 200436)

摘 要:轨道交通通信系统中的集中告警系统不判定每个子系统故障数据是否正确,也不判定某一子系统故障与另一子系统故障的关联性,对系统故障的智能识别不起作用。为了实现对系统故障的智能识别,在对 Apriori 算法的最小支持度和最小置信度进行改进的基础上,实现了动态生成最小支持度和最小置信度;对子系统的故障和子系统间的故障进行分析,挖掘其相互关联关系,提前识别出系统故障。在地铁项目上的使用表明,改进的算法可以有效识别出子系统的部分故障,以及子系统间的故障关联关系。

关键词:数据挖掘;智能;告警;故障关联

中图分类号:TP311.1 **文献标识码**:A **文章编号**:1672-755X(2021)02-0007-05

Research on the Application of the Improved Apriori Algorithm in Centralized Alarm System

LI Wen-feng[1], YAN Tao[2]

(1. Jinling Institute of Technology, Nanjing 211169, China; 2. China Railway Signal & Communication Shanghai Engineering Bureau Group Co., Ltd., Shanghai 200436, China)

Abstract: The centralized alarm system widely used in the rail transit communication system does not determine whether the fault data of each subsystem is correct, nor does it determine whether the fault of one subsystem is related to another. It has no effect on the intelligent identification of system faults. In order to realize the intelligent identification of system faults, based on the improvement of the minimum support and minimum confidence of the Apriori algorithm, the minimum support and minimum confidence are dynamically generated, and the faults of subsystems or the faults between subsystems are analyzed. The research tries to mine their interrelationships to identify system faults in advance. Through the use of actual subway projects, the results show that the improved algorithm can effectively identify partial faults of the subsystems, and can also effectively identify the relationship between the faults of the subsystems.

Key words: data mining; intelligence; alarm; fault interrelationships

随着我国轨道交通的快速建设,大量的系统设备已经步入运维管理阶段。然而,由于我国前期重点关注轨道交通系统的建设及建设人才的培养,导致轨道交通行业在检修运维技术尤其是故障诊断和寿命预测方面的能力较差[1]。基于传统的故障维修、计划维修手段开展故障判断及处理时,过度依赖人工经验,无法为后续的运维提供有效的数据支持和智能化指导[2]。目前,集中告警系统只是负责数据的采集、处理

收稿日期:2020-12-01
基金项目:金陵科技学院高层次人才科研启动基金(jit-b-202109)
作者简介:李文锋(1976—),男,江西于都人,正高级工程师,博士,主要从事数据挖掘、智能运维研究。

等基本操作，无法判定每个子系统的故障数据是否正确，也无法判定某子系统的故障是否由另一子系统故障导致。每个子系统之间虽然在功能上相互独立，但是出现故障时有可能相互关联，比如，网络故障将导致多个子系统同时出现故障。如何处理故障并利用相关数据分析预测设备状态和故障，目前还没有成熟的解决方案[3-6]。随着轨道交通业务日益发展，设备智能运维问题逐渐受到重视。目前此类研究工作主要是从综合监控系统角度提出某一具体系统内部设备故障的关联性分析。文献[7－8]主要利用专家知识库对发生的故障类型进行判定并提供处理措施供值班员参考。文献[9]提出构建故障属性集建立样本空间，根据当前监测数据，推断可能跟此故障相关的新故障，供值班人员尽早发现风险隐患，降低运营风险。但是，有关系统与系统之间故障相互关联的研究成果尚未见报道。目前基于数据挖掘的主要技术有决策树、神经网络、回归、关联规则等。本文试图利用关联规则分析 Apriori 算法[10]对轨道交通专用通信集中告警系统收集的子系统故障的关联关系进行挖掘，找出子系统之间故障的关联关系，挖掘系统存在的故障隐患。

1 Apriori 算法应用分析

对于商品推荐或其他类似的应用领域，商品本身代表一个单一样本，如商品中的啤酒、尿布等，它们可能被人为布置在同一次交易记录中，则认为此项目中的物品相互关联，符合采购人的采购意愿。利用 Apriori 算法可以挖掘出物品间的关联性。对于轨道交通专用通信系统而言，所有故障都隶属于不同子系统，虽然不是人为控制某子系统发生故障，但是故障的出现与特定的使用人群、使用环境是有关系的。因此，利用 Apriori 算法分析轨道交通通信系统的故障是可行的。

在实际的构建过程中，Apriori 算法的支持度与置信度均是设定的阈值。对于 Apriori 算法在轨道交通中的应用，仅靠人为设定支持度与置换度阈值不能满足应用要求。轨道交通集中告警系统在分析通信子系统间的故障时，不仅仅要找出经常出现的组合项，还要找出事务中不出现的项目数。Apriori 算法应用于轨道交通的特点如下：1)单系统的故障数小。单系统的故障数小说明系统的性能相对稳定，出现故障的次数少。但是，随着时间的推移，任何设备或系统的故障数都会增加，因此可以根据故障数小推断系统可能会出现异常。2)多系统间故障具有关联性。多系统出现故障的组合数体现系统间故障发生的关联性，组合数越多，说明系统间同时出现故障的次数越多。因此，利用多系统间故障的组合数可以判定系统间故障的关联性。

基于以上讨论，为了使 Apriori 算法更好地应用于轨道交通专用通信系统间的故障挖掘，本文对 Apriori 算法的关联规则进行改进，以满足轨道交通通信系统间故障挖掘的要求。

2 改进的 Apriori 算法的描述

2.1 相关术语

项目：最小的表示单位。一个事务中可能包含一个或多个项目。$T_i=\{I_1,I_2,I_3,\cdots,I_k,\cdots,I_n\}$，$T_i$ 表示第 i 个事务，I_k 表示第 $k(k\leqslant n)$ 个项目，共有 n 个项目。

项集：由多个事务组成的集合。项集 $D_i=\{T_1,T_2,T_3,\cdots,T_i,\cdots,T_m\}$，共有 m 个事务。

子数据集：满足一定条件的事务的集合。$D_{s,x}=\{T_O,\cdots,T_P,\cdots,T_Q,\cdots\}$，其中，$O,P,Q$ 表示事务在项集中的序号，$O,P,Q\leqslant m$，$D_{s,x}$ 表示第 x 个子数据集。

2.2 算法思路

改进的 Apriori 算法(以下简称“改进算法”)是在 Apriori 算法的基础上，对预先设定关联规则中最小支持度和最小置信度的方式进行改进。将项集划分为不同的事务和子数据进行构建，根据给定的事务计算子数据集的支持度和置信度，然后计算项集的最小支持度的平均值并将其作为最小支持度，计算项集的最小置信度的平均值并将其作为最小置信度。具体描述如下：

1)构建事务。根据项目出现的时间，以固定时间范围为单位，将固定时间范围内的所有项目构成一个事务，以同样的方式构建其他事务。

2）建立子数据集。根据事务发生的时间顺序，以固定的、连续的时间段为单位，构建子数据集。在建立子数据集时，如果出现某个事务与前后相邻的事务不连续，则为其单独建立一个子数据集。

3）扫描各子数据集事务中各项目的出现次数，计算当前子数据集中的最小置信度和最小支持度，找出对应的一元频集；将各子数据集的最小支持度和最小置信度分别求均值，计算出子数据集的最小支持度和最小置信度。

4）产生候选项集。前一频繁集与自身连接产生当前候选集。

5）计算项集的最小支持度和最小置信度，产生项集的频繁集。

2.3　动态关联规则

2.3.1　最小支持度

在候选集确定后，根据最小支持度筛选项目到频繁集中。每次采取固定的最小支持度会产生大量无用项集。为了避免 Apriori 算法采用固定的最小支持度带来问题，可以先计算任意 2 个项目的支持度，然后计算所有事务中项目的平均支持度，以此支持度作为最小支持度进行候选集的构建。

假定项目 $I_i \rightarrow I_j$ 的支持度用 $S_i(I_i \rightarrow I_j)$ 表示，表示“同时出现 I_i, I_j 的事务次数”与“总的事务次数”的比：

$$S_i(I_i \rightarrow I_j) = \frac{\sigma_i(I_i \bigcup I_j)}{N_i} \tag{1}$$

其中，$\sigma_i(I_i \bigcup I_j)$ 表示项集中 I_i, I_j 同时出现的事务次数，N_i 表示总的事务次数。对于项集任意的项目 I_i 出现的次数，可以表示为：

$$\sigma_i(I_i) = |T_i \mid I_i \subseteq T_i, T_i \in T| \tag{2}$$

其中，T_i 表示某个事务，T 表示事务的集合。

则对于任意项目数量大于 2 的项集的支持度 $S_i(I_i \rightarrow I_j)$，表示为 N 次事务中支持度 $S_i(I_i \rightarrow I_j)$ 的平均值。

$$S(I_i \rightarrow I_j) = \sum_{i=1}^{N} S_i(I_i \rightarrow I_j)/N \tag{3}$$

其中，$S_i(I_i \rightarrow I_j)$ 表示同时出现 I_i, I_j 的事务的支持度，将此平均值作为最小支持度，见公式(4)。

$$S'(I_i \rightarrow I_j) = \frac{\sigma_i(I_i \bigcup I_j)}{N} \tag{4}$$

如果项目数量为 1，则项集的支持度设定初值为 1。

2.3.2　最小置信度

在候选集确定后，根据最小置信度筛选项目到频繁集中。每次采取固定的最小置信度同样会产生大量无用项集。同理，可以先计算任意 2 个项目的置信度，然后计算所有事务中项目的平均置信度，以此置信度作为最小置信度进行候选集的构建。

任意项目数量大于 2 的项集的置信度 $C(I_i \rightarrow I_j)$ 定义为 N 次事务中所有最小置信度的平均值。

$$C(I_i \rightarrow I_j) = \sum_{i=1}^{N} C_i(I_i \rightarrow I_j)/N \tag{5}$$

其中，$C_i(I_i \rightarrow I_j)$ 表示同时出现 I_i, I_j 的事务的置信度，计算公式见式(6)。

$$C_i(I_i \rightarrow I_j) = \frac{\sigma_i(I_i \bigcup I_j)}{\sigma_i(I_i)} \tag{6}$$

如果项目数量为 1，则项集的置信度设定初值为 1。

2.3.3　差集

对于项集 D_i, D_j，所有出现在项集 D_i 而不出现在项集 D_j 的项目 I_i 的集合，称为差集 P_i，表示为：

$$P_i = \{I_i \mid I_i \in D_i, I_i \notin D_j\} \tag{7}$$

2.4　算法描述

算法实现过程主要是将项目归类到事务中，根据事务计算平均支持度及置信度，最后计算候选集及频繁集，具体伪代码如下：

输入：D－项目数据库
输出：所有的频繁集
For(i=1；i<=项目总数；i++){
If(D_{i+1}的时间属性－D_i的时间属性<=常数){
将 D_{i+1} 放入事务 T_j 中；
}
Else{
将 D_{i+1} 放入事务 T_{j+1} 中；
}
}
For(k=1；k<=事务总数；k++){
利用最小支持度及最小置信度构建候选集，并计算出频繁集 L_k；
}
Return L_k=所有的频繁集

3 改进的 Apriori 算法的性能分析

为了验证改进算法的效果，在配置八核至强处理器 E7、装有 Windows Server 2008 操作系统的服务器上进行以下试验。将 Apriori 算法和改进算法（以下图中标识为 TApriori 算法）在相同的测试数据集上产生频繁集时的准确率进行比较。Apriori 算法在最小支持度和最小置信度分别预先设定为 10％的情况下进行验证，本文改进算法则根据实际数据情况动态产生的最小支持度和最小置信度得到各项目频繁集。如图 1 所示，产生一元和二元频繁集时，两种算法产生的频繁集组合数一样；产生三元、四元、五元频繁集时，本文改进算法相比 Apriori 算法减少了频繁集组合的数量，可以更快地获得频繁集。

在服务器上分别运行两种算法，在不同的事务规则下测试两种算法的运行时间。分别假定事务数为 5、10、100、1 000 和 1 500 个，每个事务都有 8 个元素。试验结果如图 2 所示。试验结果表明，在事务数量较小的情况下，改进算法的运行时间与 Apriori 算法的运行时间相差不大；随着事务数量的增多，改进算法的运行时间比 Apriori 算法的运行时间明显减少。

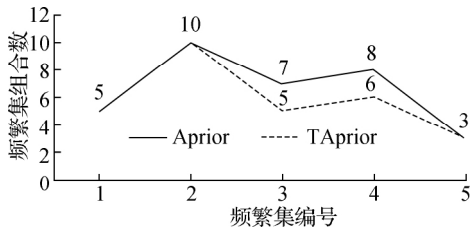

图 1 两种算法产生的频繁集组合数比较

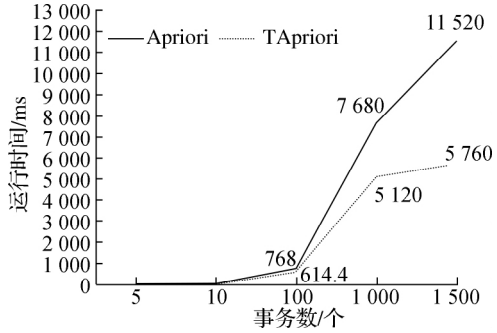

图 2 不同事务数对应的运行时间

4 改进的 Apriori 算法的应用分析

4.1 在 A 线路上的应用分析

将改进算法用于轨道交通专用通信集中告警系统上，对实际应用的某城市轨道交通 A 线路 2019 年 1 月—2020 年 3 月共 417 737 次告警数据进行分析。为了便于统计，设定各子系统代码集如表 1 所示。

利用改进算法进行故障关联挖掘，A 线路分别生成二元频繁集 L2、三元频繁集 L3、四元频繁集 L4。每种频繁集反映不同子系统同时出现故障的可能性，即故障的关联性。二元、三元和四元频繁集分别体现 2 个、3 个和 4 个子系统之间的故障关联关系。

表 1　子系统代码表

子系统名称	子系统代码	子系统名称	子系统代码
时钟系统	1	计算机网络系统	6
专用电话系统	2	电源系统	7
乘客系统	3	专用无线系统	8
公务电话系统	4	传输系统	9
广播系统	5	视频监控系统	10

由表 2 可知，A 线路形成的二元频繁集 L2 中，子系统 3、5 同时出现故障的次数最高，达到 194 次；子系统 3、7 同时出现故障的次数次之，达到 172 次；接下来子系统 5、7 和子系统 2、3 同时出现故障的次数较高。

由表 3 可知，A 线路形成的三元频繁集 L3 中，子系统 3、5、7 同时出现故障的次数最高，达到 135 次；子系统 2、3、5 同时出现故障的次数次之，达到 101 次；接下来子系统 2、3、7 和子系统 3、5、8 同时出现故障的次数较高。

由表 4 可知，A 线路形成的四元频繁集 L4 中，子系统 2、3、5、7 同时出现故障的次数最高，达到 67 次；子系统 3、5、7、8 同时出现故障的次数次之，达到 56 次；接下来子系统 2、3、5、8 同时出现故障的次数较高。

通过分析 A 线路数据形成的 L2、L3、L4，2、3、5、7、8 项为高频项。对故障的实际明细进行排查，其中一座车站被改造为换乘站，分布在各站的专用电话系统、乘客系统、广播系统、电源系统、专用无线系统及其设备受网络不稳定影响，频繁出现故障。表 2—表 5 按故障次数从小到大的顺序排列。

表 2　A 线路形成的二元频繁集 L2

子系统代码	同时出现故障的次数
1、3	74
2、3	123
2、5	102
2、7	83
3、5	194
3、7	172
3、8	94
5、7	135
5、8	81

注：同时出现故障次数较少的子系统组合未在表中列出，下同。

表 3　A 线路形成的三元频繁集 L3

子系统代码	同时出现故障的次数
2、3、5	101
2、3、7	83
3、5、7	135
3、5、8	81

表 4　A 线路形成的四元频繁集 L4

子系统代码	同时出现故障的次数
2、3、5、7	67
2、3、5、8	43
3、5、7、8	56

4.2　在 B 线路上的应用分析

通过对实际应用的某城市轨道交通 B 线路 2017 年 12 月—2020 年 3 月的共 1 299 次告警情况进行分析得到各子系统的故障总数，如表 5 所示。

表 5　B 线路各子系统故障总数

子系统代码	故障总数
1	14
2	66
4	52
5	282
7	208
8	400
9	30
10	0

通过表 5 可以看出，B 线路集中告警系统近 3 年收到子系统代码为 10 的视频监控系统的故障数为 0，与实际情况不符。前面提到单故障数小说明系统的性能相对稳定，出现故障的次数少。但是，随着时间的推移，任何设备或系统的故障将会变多，故障数应该增加。可以推断视频监控系统有异常。后期经过值班人员排查，发现人为将视频监控系统的故障上报功能屏蔽了，所有该子系统的告警都不上报给集中告警系统。

5　结　语

在集中告警系统中引入改进的 Apriori 算法后，不仅完成了集中告警系统具备的故障采集、（下转第 32 页）

找到较合适的模型参数，这与传统的通过数据实验确定模型参数的结论一致。这里利用梯度下降法对参数调整方法进行了详细说明，并对参数调整的顺序和多层卷积作了说明，为基于多层卷积的复杂网络模型的有效性证明奠定基础。由于复杂卷积神经网络模型的结构与基本卷积模型类似，只要损失函数 E 是输入 X^n 的凸函数，本文的证明就依然适用。但是复杂卷积神经网络的损失函数 E 经过多层运算后，通常不再是输入 X^n 的凸函数，此时需要利用优化方法的有关数学理论将损失函数转换成 X^n 的凸函数，而如何将非凸函数转化成凸函数有待进一步研究。

参考文献：

[1] LECUN Y，BOSER B，DENKER J S，et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation，1989，11(4)：541－551

[2] LECUN Y，BOTTOU I，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278－2324

[3] KRIZHEVSKY A，SUTSKEVER I，HINTON G. ImageNet classification with deep convolutional neural networks[C]. Cambridge MA：MIT Press，2012：1106－1114

[4] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston：IEEE，2015：1－9

[5] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas：IEEE，2016：770－778

[6] MCNEELY-WHITE D，BEVERIDGE J R，DRAPER B A. Inception and ResNet：same training，same features[J]. Springer Cham，2019，48：352－357

[7] MCNEELY-WHITE D，BEVERIDGE J R，DRAPER B A. Inception and ResNet features are (almost) equivalent[J]. Cognitive Systems Research，2020，59：312－318

[8] VIDAL R，BRUNA J，GIRYES R，et al. Mathematics of deep learning[J]. Machine Learning，2017，arXiv：1712.04741

[9] SNOEK J，LAROCHELLE H，ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]. Lake Tahoe：Curran Associates Inc，2012：2951－2959

（责任编辑：湛 江）

（上接第11页）处理、存储、查询、显示等功能，还实现了对各子系统故障数据的正确解析和挖掘，而且可以判定子系统与子系统之间故障的关联关系。在线路上的实际应用结果表明，引入改进算法后可以有效识别单个子系统的异常情况，并进一步挖掘出子系统间的关联故障，为轨道交通专用通信系统的运维提供支持。

参考文献：

[1] 彭飞，张尧. 轨道交通轴承故障诊断与寿命预测技术综述[J]. 城市轨道交通研究，2020，23(12)：162－168

[2] 段亚美，施聪，黄晓荣. 基于故障预测与健康管理技术的城市轨道交通信号系统健康管理体系[J]. 城市轨道交通研究，2020，23(12)：177－181

[3] 周勇. 城市轨道交通智慧车站技术方案研究与实现[J]. 铁道建筑，2020，60(12)：117－120

[4] 刘彦军，杨涛存，武威，等. 基于大数据技术的高铁运营安全规律分析系统设计与应用[J]. 中国铁路，2020(9)：28－33

[5] 刘丙林，朱佳，李翔宇. 城市轨道交通车辆智能运维系统探索与研究[J]. 现代城市轨道交通，2019(6)：16－21

[6] 白华. 城市轨道交通车辆智能运维系统研究[J]. 科学与技术，2020(35)：1－2

[7] 杨莎，张振山. 基于关联关系的城市轨道交通智能报警系统[J]. 城市轨道交通研究，2017，20(12)：70－72

[8] 刘雅娟. 基于故障诊断与健康管理专家系统的装备体系架构[J]. 无线电通信技术，2021，47(3)：259－268

[9] 张铭，王富章，程超. 城市轨道交通设备故障聚类与贝叶斯网络预警[J]. 计算机工程与应用，2016，52(11)：259－264

[10] AGRAWAL R，IMICLINSKI T，SWAMI A. Mining association rules between sets of items in large database[C]. Washington D C：Proceeding of the ACM SIG-MOD Conference on Management of Data(SIGMOD 93)，1993：207－216

（责任编辑：湛 江）