文章编号：1001-4632（2021）04-0155-11

基于优化的MsEclat算法的铁路机车事故故障关联规则挖掘

李 鑫[1]，史天运[2]，常 宝[3]，马小宁[3]，刘 军[3]

（1. 中国铁道科学研究院 研究生部，北京 100081；
2. 中国铁道科学研究院集团有限公司 科技和信息化部，北京 100081；
3. 中国铁道科学研究院集团有限公司 电子计算技术研究所，北京 100081）

摘 要：为从铁路机车大数据中挖掘出与机车事故故障有关的关联规则，提出1种优化的MsEclat算法。先提出改进的Eclat算法——MsEclat算法，构建最小支持度索引表，以各项目的支持度值为排序依据重新构建数据集，依据垂直挖掘思想获得针对不同项目的频繁项集，解决Eclat算法无法在多最小支持度情况下挖掘关联规则的缺陷；进一步改进得到优化的MsEclat算法，在融合布尔矩阵、并行计算编程模型MapReduce基础上，设计频繁项集挖掘步骤，提高算法在大数据分析场景下的执行效率。通过算法对比，验证MsEclat算法及其优化算法在多最小支持度关联规则挖掘方面的计算效率优势。最后，以某铁路局的机车运转养护大数据为例，采用优化的MsEclat算法，挖掘机车事故故障的关联规则。结果表明：该算法在6个分布式节点的情况下耗时3.945 034 s，挖掘得到频繁项集156条，如运用故障高发的机车中，83.78%的概率会同时出现频次较多的行车安全装备问题等；形成相应关联规则后，可用于分析该局机车的事故故障发生情况及质量安全状态。

关键词：机车事故故障；关联规则；大数据分析；数据挖掘技术；MsEclat算法；多最小支持度
中图分类号：TP301.6 **文献标识码**：A **doi**：10.3969/j.issn.1001-4632.2021.04.18

铁路机车的设备安全、周转效率和维修质量等对于保障铁路运输生产效率有着重要的作用。随着机车数量的增加、机车数据的不断积累和数据获取方式的不断升级[1]，机车积累的运行台账、监测检测、维修保养、事故故障等各类数据不断增多[2]，数据类型日益丰富，蕴含的数据价值不断显现。通过研究机车各类数据间存在的关联关系，可以挖掘出机车事故故障与机车运用、整备、检修等日常生产数据之间存在的关联规则，这将对提高机车运行质量、提升故障诊断效率和安全管理能力、促进机车检修作业精细化水平起到较好的促进作用。然而，由于数据挖掘技术在铁路机务专业起步较晚，目前仍较难从大量数据中发掘蕴含的重要信息，特别是与机车事故故障有关的关联性信息。因此，运用数据挖掘技术中的关联规则分析方法，对机车事故故障及其关联因素的探究，一直是铁路机车质量安全管理中的重点和难点。

关联规则挖掘是数据挖掘中的重要组成，其主要任务是挖掘出频繁项集并生成关联规则[3]。在实际的铁路机车运输生产中，事故、故障、设备问题等数据相较于机车的日常运用、整备、检修数据有着较低的发生概率，但是其对机车的质量及安全状态却有着重要的影响。在对机车事故故障进行关联规则挖掘时，若以上所有数据均采用统一的支持度，很容易因设定的支持度过高而无法获得与机车事故故障有关的关联规则，或因设定的支持度过低而产生过多低价值的关联规则。因此，进行铁路机车事故故障关联规则挖掘时，需要对具有不同发生概率及重要性的数据设定不同的支持度，从而获得更有价值的关联规则。

目前最常用的多最小支持度关联规则挖掘算法主要有2类。一类是基于先验算法（Apriori Algorithm）形成的多最小支持度先验算法（MsApriori Algorithm）[4]及其相关改进算法，这类算法延续

收稿日期：2021-01-13；修订日期：2021-03-31
基金项目：中国国家铁路集团有限公司科技研究开发计划重大课题（P2019G003）
第一作者：李 鑫（1990—），男，山西闻喜人，博士研究生。E-mail：florian_lee@163.com
通讯作者：史天运（1967—），男，山西平陆人，研究员，博士研究生导师，博士。E-mail：shitianyun@sina.com

了Apriori算法的基本思路，在数据挖掘过程中通过多次扫描数据库，计算对应候选项集的支持度，进而筛选出频繁项集[5]。但这个过程在增加内存负担的同时还会产生大量的候选项集[6]，存在数据集搜索次数多、内存开销大、时间损耗严重等问题。另一类是利用频繁模式增长算法（FP-Growth Algorithm）[7]的思想，在树结构[8]的基础上实现多最小支持度的关联规则挖掘，如多最小支持度频繁模式增长算法（MS-Growth Algorithm）[9]等，这类算法虽然在执行速率上较MsApriori算法有一定提升，但是普遍对内存要求较高，且在挖掘的过程中需多次扫描构建的树，面对较大的数据集时会造成算法效率快速下降[10]。这2类算法均属于水平挖掘算法[11]。

相对于水平挖掘算法，还有以垂直方式进行挖掘的关联规则挖掘算法，这种算法的效率较高，但无法实现多最小支持度情况下的关联规则挖掘。例如等价变换类算法（Eclat Algorithm）[12]，只需扫描1次数据库，即可快速求得候选项集的支持度，计算性能优于Apriori算法等水平挖掘算法[13]，但尚不能直接应用于机车事故故障的关联规则挖掘场景。

本文基于运用深度优先策略的Eclat算法思路，通过构建最小支持度索引表，以各项目的最小支持度值为排序依据重新构建垂直格式数据集，形成基于多最小支持度的Eclat改进算法——MsEclat算法，使其能够在多最小支持度情况下挖掘出关联规则；再对MsEclat算法进一步改进，将算法与位运算求交集、等价类并行运算这2种数据处理方法加以融合，形成优化的MsEclat算法，使其更好地应用于大数据分析场景，能够更加高效地挖掘出与机车事故故障等小概率项目有关的频繁项集；通过算法对比，验证MsEclat算法及其优化算法在多最小支持度关联规则挖掘方面的计算效率优势；以某铁路局2019—2020年的机车运转养护大数据为例，采用优化的MsEclat算法，挖掘机车事故故障的关联规则，再进一步以筛选得到的6条代表性关联规则为例，分析该局机车的事故故障发生情况及质量安全状态。

1 MsEclat算法的背景知识

1.1 垂直格式数据集

垂直事务集合T_{set}。记事务集合为T=$\{t_1,t_2,\cdots,t_n\}$，项目集合为$I=\{i_1,i_2,\cdots,i_m\}$，若T与I之间存在一定的对应关系，二者共同构成包含事务与项目的数据集，那么该数据集有2种排序表示方式。若按事务对数据集排序，形成的数据集可称之为水平格式数据集；若按项目对数据集排序，形成的数据集则可称之为垂直格式数据集。水平格式数据集与垂直格式数据集的表示方式示例见表1和表2。对于项目集合I中的任意项目$i_j\in I$，$1\leqslant j\leqslant m$，记垂直事务集合为$T_{set}(i_j)$，表示垂直格式排序下，事务集合T中包含项目i_j的所有事务的集合。

表1 水平格式数据集示例

事务	项目集合
1	A,B,D
2	A,B,C,D
3	A,C
4	A,C,D
5	A,B,C
6	B,C

表2 垂直格式数据集示例

项目	垂直事务集合
A	1,2,3,4,5
B	1,2,5,6
C	2,3,4,5,6
D	1,2,4

1.2 支持度、置信度与提升度

1）支持度S_{sup}

引入支持度的概念，用S_{sup}表示同时包含1个或多个项目（或项目集合）的事务在事务集合中的占比。在Eclat算法中，项目集合$\{I_a\cup I_b\}(I_a\subseteq I,I_b\subseteq I)$的支持度为$S_{sup}(I_a\cup I_b)$，表示同时包含$I_a$和$I_b$的事务在事务集合$F$中的占比，其计算式为

$$S_{sup}(I_a\cup I_b)=\frac{\left|T_{set}(I_a)\cap T_{set}(I_b)\right|}{|F|} \tag{1}$$

支持度是衡量某一项目集合能否得到保留并生成关联规则的重要参数，其衡量的标准被称为最小支持度S_{min_sup}，一般可人为设定。所有项目集合在参与衡量前被称为候选项集，而筛选后未被舍弃获得保留的项目集合可被称为频繁项集。某1个含有k个项目的项目集合在进行保留或舍弃的选择前，称为候选k项集，若其支持度大于或等于S_{min_sup}，便被称为频繁k项集[14]，得到保留。

2）支持度计数值C_{sup}

项目集合的支持度与其对应的T_{set}所含事务个数密

切相关，因此引入支持度计数值的概念，用来表征项目集合的发生概率情况。设项目集合$\{I_a \cup I_b\}$ $(I_a \subseteq I, I_b \subseteq I)$的垂直事务集合为$T_{set}(I_a \cup I_b)$，则项目集合$\{I_a \cup I_b\}$的支持度计数值$C_{sup}(I_a \cup I_b)$可表示为$T_{set}(I_a \cup I_b)$中的事务个数，其计算式为

$$C_{sup}(I_a \cup I_b) = \left| T_{set}(I_a \cup I_b) \right| = \left| T_{set}(I_a) \cap T_{set}(I_b) \right| \tag{2}$$

与依据支持度衡量某1个项目集合能否得到保留的过程类似，在利用垂直格式数据集挖掘频繁项集的过程中，可采用支持度计数值的大小作为项目集合是否能得到保留的依据。同样需要人为设定1个最小支持度计数值C_{min_sup}，若某1个候选k项集的支持度计数值大于或等于C_{min_sup}，该项目集合才能得到保留并被称为频繁k项集。

3）置信度S_{con}

引入置信度的概念，用来表征项目集合之间的影响程度。对于1个关联规则$I_a \Rightarrow I_b$，置信度$S_{con}(I_a \Rightarrow I_b)$表示包含$I_a$的事务中同时也包含$I_b$的事务的比例，即出现$I_a$时也同时出现$I_b$的概率，其计算式为

$$S_{con}(I_a \Rightarrow I_b) = \frac{\left| T_{set}(I_a) \cap T_{set}(I_b) \right|}{\left| T_{set}(I_a) \right|} \tag{3}$$

4）提升度S_{lif}

置信度$S_{con}(I_a \Rightarrow I_b)$并不能完全反映$I_a$与$I_b$的相关联程度，有时会出现$S_{con}(I_a \Rightarrow I_b)$小于包含$I_b$的事务在事务集合中占比的特殊情况，这说明$I_a$出现时$I_b$也出现的概率要小于$I_b$单独发生的概率，实则表明$I_a$与$I_b$之间是相互排斥的。因此引入提升度的概念，以便更加全面地分析I_a与I_b的关系。

对于1个关联规则$I_a \Rightarrow I_b$，提升度$S_{lif}(I_a, I_b)$表示包含I_a的事务中同时也包含I_b的事务的比例与包含I_b的事务在事务集合中所占比例的比值，即“关联规则$I_a \Rightarrow I_b$的置信度$S_{con}(I_a \Rightarrow I_b)$”除以“包含$I_b$的事务在事务集合$F$中的占比”，其计算式为

$$S_{lif}(I_a, I_b) = \frac{S_{con}(I_a \Rightarrow I_b)}{\left| T_{set}(I_b) \right| / |F|} = \frac{\left| T_{set}(I_a) \cap T_{set}(I_b) \right| |F|}{\left| T_{set}(I_a) \right| \left| T_{set}(I_b) \right|} \tag{4}$$

提升度$S_{lif}(I_a, I_b)$反映了关联规则$I_a \Rightarrow I_b$中I_a与I_b的关联程度，其值大于1表明二者呈正相关性，小于1表明二者相互排斥，等于1表明二者之间没有关联性。

1.3　多最小支持度下的频繁项集判定

项目集合的最小支持度。在多最小支持度的关联规则挖掘中，以项目集合中各个项目最小支持度的最小值作为整个项目集合的最小支持度[15]。例如：若项目集合I中各项目的最小支持度分别为$S_{min_sup}(i_1)$，$S_{min\_sup}(i_2), \cdots, S_{min\_sup}(i_m)$，则$I$的最小支持度的计算式为

$$S_{min\_sup}(I) = \min\{S_{min\_sup}(i_1), S_{min\_sup}(i_2), \cdots, S_{min\_sup}(i_m)\} \tag{5}$$

1.4　面向有序项集的最小支持度索引表

最小支持度索引表。以最小支持度值（或最小支持度计数值）有序递增的方式存储所有项目的数据表。例如：若有1个项目集合$\{A, B, C, D, E\}$，各项目的最小支持度计数值分别为3，5，2，4和6，则各项目将依C，A，D，B和E的顺序存入最小支持度索引表。

2　MsEclat算法

2.1　Eclat算法简述

1）Eclat算法基本思想

Eclat算法是1种深度优先算法，利用概念格理论将垂直格式数据集划分为不同的等价类[16]，并通过T_{set}间的交集运算筛选出频繁项集。

概念格理论[17]引入等价类的概念，使数据集在各个独立的子空间内自底向上完成频繁项集的挖掘。具体来说，设U是1个由多个互不相同的集合构成的空间，如果U中任意2个集合的交集或并集也均包含于U，则U为1个概念格；若选取U中的部分集合构成空间W，且W中任意2个集合的交集或并集也均包含于W，则W为U的1个子概念格。进一步地，设1个概念格（或子概念格）中有多个由q个项目组成的有序集合，若这些集合具有相同的前缀，即前$q-1$个项目相同，而第q项不同，则这些集合属于同1个等价类。需强调的是，单独1个集合不能成为1个等价类。

基于概念格理论，可以通过对2个项目集合进行可连接性判定，从而避免项目集合间不必要的连接操作。设项目集合I_u和I_v属于同1个等价类，且均为频繁k项集，则项目集合I_u和I_v是可连接的[18]，可记为$I_u \bowtie I_v = I_u \cup I_v = \{i_{u_1}, i_{u_2}, \cdots, i_{u_(k-1)}, i_{u_k}, i_{v_k}\}$。若将每个项目集合及其中的项目按一定顺序排列，如果其中的2个频繁k项集I_u和I_v不能连接，则I_u

和I_v之后的所有项目集合都不满足连接条件，无须再次进行连接判断。

2）Eclat算法流程

在挖掘频繁项集时，首先扫描1次原始数据集，将水平格式的数据变换成垂直格式的数据；记$k=1$，k表示候选项集（或频繁项集）中项目的个数，从第1个项目开始，从上向下通过项目集合间的并集操作得到候选项集；通过对项目集合所对应的T_{set}进行交集运算并计算其中的元素个数，得到候选项集的垂直事务集合及其支持度计数值。然后将候选项集的支持度计数值与设定的最小支持度计数值比较，剔除不符合要求的项目集合，从而得到频繁项集。循环重复上述过程，以频繁k项集来产生候选$k+1$项集，直到不再有新的频繁项集产生为止。

3）Eclat算法的不足

显然，在Eclat算法下的某些具体分析场景中，不同的项目并不是均匀分布的，且重要性各不相同，需针对不同的项目设定不同的最小支持度，但是Eclat算法无法在多最小支持度的情况下挖掘出频繁项集，需要对其算法进行改进。

2.2 改进的Eclat算法——MsEclat算法

针对Eclat算法的上述缺陷，提出1种基于多最小支持度的Eclat改进算法——MsEclat算法。该算法的思路是：利用Eclat算法的基本思想，运用多最小支持度关联规则挖掘的相关理论知识，在最小支持度索引表的基础上构建新型垂直格式数据集、有序项集和等价类，在项目集合中各项目具有不同最小支持度值的情况下，有效挖掘出相关的频繁项集。

1）MsEclat算法的频繁项集挖掘步骤

第1步：扫描1次数据库后，将原始数据集由水平格式转变为垂直格式，同时将各项目（即候选1项集）按照最小支持度索引表的顺序由上到下排列。

第2步：选定目标项目X，将目标项目X及其$T_{set}(X)$设定为垂直格式数据集的首位；依序排列最小支持度索引表中排在项目X之后的项目，排项目X之前的项目则依据最小支持度定义暂不考虑；将项目X的最小支持度计数值$C_{min_sup}(X)$设定为以项目X为首项的各阶项目集合的最小支持度计数值，若$T_{set}(X)$计数值小于$C_{min_sup}(X)$，则结束分析，未挖掘到以项目X为分析目标的频繁项集。

第3步：以项目X为首项，与后续剩余项目依次连接，构成候选2项集，并将所有候选2项集的T_{set}的计数值与$C_{min_sup}(X)$做比较，剔除小于$C_{min_sup}(X)$的项目集合，由此得到以项目X为分析目标的频繁2项集；同样地，在频繁2项集的基础上按此方式继续形成频繁3项集。

第4步：以项目X为首项，依次在不同的等价类中分别完成项目集合之间的连接和支持度计数值的比较，生成高阶的频繁项集。

第5步：不断重复第4步，直到获得以项目X为分析目标的所有高阶频繁项集。

2）MsEclat算法实例

以表1中所列的数据集为例，采用MsEclat算法进行频繁项集挖掘的实例说明如图1所示。先对数据集构建相应的最小支持度索引表，再分别以不同的项目为分析目标，按频繁项集挖掘过程逐一运算，得到所有项目的高阶频繁项集。

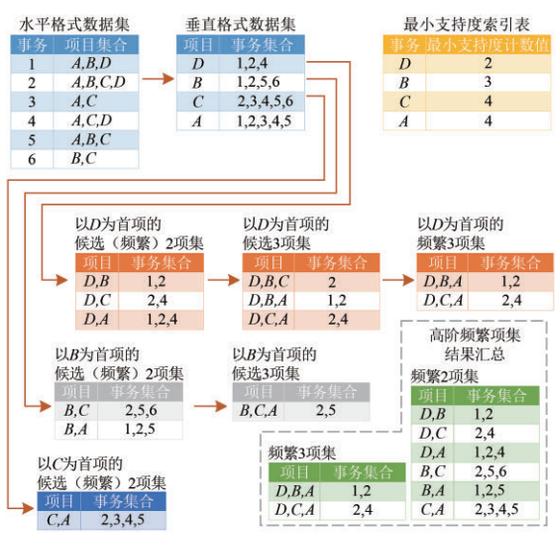

图1 采用MsEclat算法进行频繁项集挖掘的实例说明

3 优化的MsEclat算法

由MsEclat算法的计算步骤可知，常规的T_{set}间交集运算会消耗一定的时间及内存，将其应用于铁路机车大数据分析时，因数据量巨大，将会影响算法的执行效率。因此将位运算求交集、等价类并行运算这2种数据处理方法融入MsEclat算法中，通过改善T_{set}间的交集运算效率，同时利用等价类开展并行运算，使得优化的MsEclat算法在处理较大规模数据集时的数据挖掘效率进一步提高。

3.1 基于布尔矩阵的T_{set}位运算求交集

常规的T_{set}间的交集运算是从第1位开始，从

左向右依次与另1个项目集合的T_{set}进行比对，并按此循环，直到完成全部比较。这样的交集运算较为简便，但随着事务数据规模的不断增大，比较次数和比较时间会成倍增长，导致算法的挖掘效率随之降低。

引入布尔矩阵[19]，将T_{set}间常规的交集运算转变为2个1行n列矩阵的与运算。其中，n为垂直事务集合中的最大事务个数，矩阵中有事务值的对应位置设定为“1”，其余设为“0”。与运算后得到的新的矩阵，便是2个项目集合合并后新产生的T_{set}所对应的布尔矩阵。这样，当事务数据的规模很大时，只需对2个布尔矩阵进行与运算，统计新产生的矩阵中“1”的个数，即得到新项目集合的支持度计数值。这样便在很大程度上降低了算法的时间复杂度，提高了挖掘效率。

3.2　基于MapReduce的等价类并行运算

根据概念格理论，数据集（概念格）按照可连接性划分为多个等价类，并依次在各个等价类内独立地产生各阶频繁项集。当数据量比较大时，串行计算会消耗大量的时间，影响数据挖掘的效率。因此在处理大规模数据集时，可以引入并行计算编程模型MapReduce[20]，通过并行运算获得3阶及以上的频繁项集。所有的并行运算过程均抽象成映射阶段（Map）和归约阶段（Reduce）这2个处理过程，避免了考虑工作调度、容错处理、网络通信、负载平衡等细节[21]。Map阶段中各个节点读取相应的等价类，其中各个项目集合分别连接得到相应的高阶候选项集，对应的T_{set}完成交集运算。之后转入Reduce阶段，对新求得的高阶候选项集及其T_{set}完成频繁项集的筛选。最后汇总输出，得到高阶频繁项集。

3.3　大数据场景下的频繁项集挖掘步骤

大规模数据集的事务或者项目规模往往极为庞大，其所对应的候选项集和垂直事务集合的规模也十分巨大。在这样的分析场景下，还需进一步改进MsEclat算法，通过优化的MsEclat算法进一步提高运算效率，其数据挖掘过程可分为5个步骤。

第1步：初始化。

首先将各项目按照最小支持度计数值递增的顺序，构建对应的最小支持度索引表。之后扫描数据库，将水平格式数据转化为M行、$N+1$列的垂直格式数据表，其中M为项目个数，N为最大事务个数。数据表最左侧1列依最小支持度计数值递增顺序由上向下依次填入各项目名称；各项目则按照各自的T_{set}在后面N列形成相应的布尔矩阵形式的垂直事务集合。后续阶段将以新产生的垂直格式数据表为基础挖掘频繁项集。

第2步：挖掘目标项目的频繁2项集。

选定需研究的目标项目Y，以项目Y为首项，按照最小支持度索引表中的项目顺序，以第1步中新产生的垂直格式数据表为数据基础，将项目Y与其后面的项目依次连接，得到与项目Y相关的候选2项集。其中，各项目的垂直事务集合则按照位运算的方法完成交集运算得到候选2项集的垂直事务集合。以项目Y的最小支持度计数值$C_{min_sup}(Y)$为筛选条件，剔除事务数小于该值的候选项集，从而得到与项目Y相关的频繁2项集。

第3步：获得频繁3项集的等价类。

以第2步中得到的频繁2项集为基础，所有频繁2项集依次与其后面的项目集合连接，得到与目标项目Y相关的候选3项集以及对应的垂直事务集合。依然以$C_{min_sup}(Y)$为筛选条件，得到与项目Y相关的频繁3项集。之后，按照概念格理论，将频繁3项集划分为不同的等价类，作为下一阶段的输入。

第4步：得到频繁k项集。

调用并行计算编程模型MapReduce，将上一阶段得到的等价类分配到不同的节点，进行频繁k项集的挖掘。然后重新划分等价类并循环该步骤，直到获得与项目Y相关的最高阶频繁项集为止。

第5步：输出结果。

汇总各阶段得到的与项目Y相关的各阶频繁项集及其垂直事务集合，以供后续分析、计算之用。

如需得到所有项目相关的各阶频繁项集，只需依照最小支持度索引表依次选定各个项目为目标项目，不断循环第2～5步，并对结果加以汇总即可。

4　算法对比

分别开展MsEclat算法与MsApriori算法、MS-Growth算法这2种水平挖掘算法的对比，以及MsEclat算法与其优化算法的对比，从计算时间的角度考察算法效率。处理规模极大的数据集时，MsApriori算法与MS-Growth算法将无法进行有效运算，而在小规模数据集的情况下，又很难表现出MsEclat算法与其优化算法的差异，故2次对比实验选用不同规模的数据集。

4.1 MsEclat算法与水平挖掘算法对比

实验环境：处理器为 Intel i5-8250u 1.8 GHz，内存 8 GB，操作系统为 Windows 10，软件环境为 python 3.7.3，开发环境为 PyCharm。

选取数据规模和数据稠密程度各不相同的 3 个数据集 Mushroom，Pumsb_star 和 Car Evaluation（分别取自关联规则挖掘数据库 http://fimi.uantwerpen.be/data/以及 https://archive.ics.uci.edu/），根据各个数据集不同的稠密程度，将数据集中的项目按照 30%，40% 和 30% 的比例赋予不同的最小支持度，对比 MsEclat 算法与 MsApriori 算法、MS-Growth 算法挖掘频繁项集所用的时间，计算结果见表 3。

表 3 3个数据集下不同算法的计算结果

数据集	不同比例下的最小支持度			不同算法的执行时间/s			频繁项集数/条
	30%	40%	30%	MsEclat	MsApriori	MS-growth	
Mushroom	0.4	0.6	0.8	0.125 009	2 455.328 992	172.467 558	169
Pumsb_star	0.6	0.7	0.8	1.874 561	1 415.279 994	94.732 062	135
Car Evaluation	0.1	0.2	0.4	0.015 629	2.891 171	0.750 397	109

从计算结果可以看出，MsEclat 算法的时间开销明显优于同属于水平挖掘算法的 MsApriori 算法和 MS-Growth 算法。分析其原因，一是 MsEclat 算法继承了 Eclat 算法的垂直数据格式，只需遍历 1 次数据库即可进行频繁项集的挖掘工作；二是 MsEclat 算法在频繁项集的挖掘过程中，仅涉及垂直事务集合的交集运算以及最小支持度的大小比较等简单操作，时间和内存消耗不高；三是 MsEclat 算法依然延续了概念格与等价类的思想，避免了项目集合之间不必要的连接操作，进一步提高了挖掘效率。

4.2 MsEclat算法与其优化算法对比

实验环境：最大节点数为 10 的分布式计算集群（Hadoop），每个节点的处理器为 Intel i5-5200u 2.2 GHz，内存 8 GB，操作系统为 Ubuntu Linux。

由数据生成器分别产生规模约为 0.5，1.1，1.7 和 2.3 GB 的数据集，每个数据集的项目数为 1 000 个，事务数分别为 3×10^6，6×10^6，9×10^6 和 12×10^6 条。每个数据集中的项目按照 30%，40% 和 30% 的比例进行划分，并分别赋予 0.4，0.6 和 0.8 的最小支持度。节点数为 6 时，MsEclat 算法与其优化算法对不同规模数据集的处理时间如图 2 所示。由图 2 可以看出：MsEclat 算法经过 2 种数据处理方法的优化后，计算效率明显提高；随着数据规模的增大，优化算法在计算时间上的优势愈发明显。其余节点数时的实验情况与此类同，不再展开。

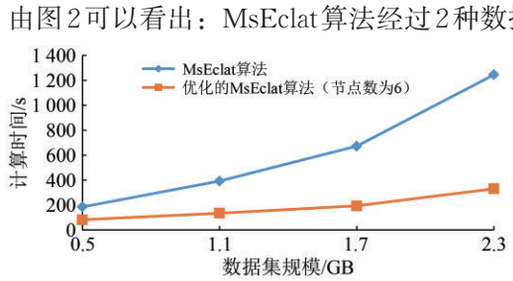

图 2 MsEclat 算法与其优化算法对不同规模数据集的处理时间

节点数不同的情况下，优化的 MsEclat 算法对不同规模数据集的处理时间如图 3 所示。由图 3 可以看出：随着计算节点的增加，优化的 MsEclat 算法对不同规模的数据集的处理时间逐渐减少；当计算节点增加到一定的数量后，受分布式计算集群间通信时间的影响，算法计算时间的减少程度逐渐减弱。

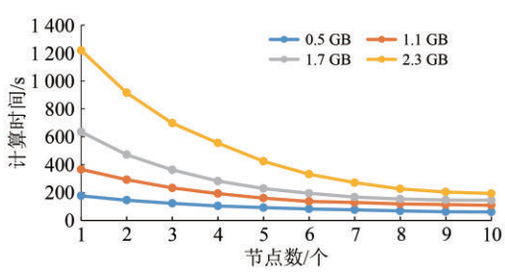

图 3 优化的 MsEclat 算法在不同节点数时对不同规模数据集的处理时间

5 算例分析

现以某铁路局 2019—2020 年的机车数据为例，采用优化的 MsEclat 算法，进行机车事故故障的关联规则挖掘。

5.1 待分析项目选取

从不同维度，选取可用于开展机车事故故障关联分析的项目共计 64 项，例如机车运用质量、整备质量、检修质量、专项整治、综合评价和基本信息等，详见表 4。根据该局 2019—2020 年机车数

表4　机车事故故障关联分析的项目

项目基本维度	项目细分维度	项目及其判定
运用质量	机破	机破发生1次 机破发生2次 机破高发(每月发生2件以上)
	运用故障	运用故障少量(每月发生3件以内) 运用故障较多(每月发生3～5件) 运用故障高发(每月发生5件以上)
	临修	临修少量(每月1次) 临修较多(每月2～3次) 临修高发(每月3次以上)
	碎修	碎修少量(每月发生10件以内) 碎修较多(每月发生10～20件) 碎修过多(每月发生21～30件) 碎修超多(每月发生30件以上)
整备质量	五项专检问题	五项专检问题少量(每月发生1～5件) 五项专检问题较多(每月发生6～10件) 五项专检问题很多(每月发生11～15件) 五项专检问题过多(每月发生15件以上)
	行车安全装备问题	行车安全装备问题较少(每月发生1～3件) 行车安全装备问题较多(每月发生4～6件) 行车安全装备问题过多(每月发生6件以上)
检修质量	零公里质量鉴定问题	零公里质量鉴定问题少量(每月发生1～3件) 零公里质量鉴定问题较多(每月发生4～6件) 零公里质量鉴定问题很多(每月发生7～12件) 零公里质量鉴定问题过多(每月发生12件以上)
	性能试验问题	性能试验问题少量(每月发生1件) 性能试验问题较多(每月发生2件) 性能试验问题过多(每月发生2件以上)
	必换件问题	必换件问题发生
	修程	和谐型机车:C1,C2,C3,C4,C5,C6
		其他类型机车:辅修,小修,中修,大修
专项整治	机车春季鉴定整修	春鉴等级优秀,春鉴等级良好,春鉴等级合格,春鉴等级不合格
	机车秋季鉴定整修	秋鉴问题发生
综合评价	总体质量评价	A级机车,B级机车,C级机车,D级机车
基本信息	支配单位	H1机务段,H2机务段,H3机务段
	生产厂家	G1机车车辆厂,G2机车车辆厂,G3机车车辆厂,G4机车车辆厂,G5机车车辆厂
	担当线路	L1线,L2线,L3线,L4线,L5线,L6线,L7线,L8线,L9线

据，得到的事务集合中共有事务31 332条，每1条事务中的项目均由上述这64个项目组成。

根据数据占比、重要程度、专家意见等因素，对机破、运用故障、碎修、五项专检问题、性能试验问题、机车春季鉴定整修、总体质量评价这些细分维度下的项目赋予不同的最小支持度计数值，并形成最小支持度索引表，见表5。表中项目的最小支持度计数值越小，表示该项目在数据集中出现频次较低，但重要程度较高。表中未列出的其他项目的最小支持度统一定为0.2，即最小支持度计数值为6 266。

5.2　关联规则挖掘

采用优化的MsEclat算法进行挖掘，在6个分布式节点的情况下耗时仅3.945 034 s，挖掘得到频繁项集156条，其中频繁1项集24条、频繁2项集60条、频繁3项集54条、频繁4项集17条、频繁5项集1条。结合业务部门意见，从中筛选出6条具有代表性的频繁项集形成相应的关联规则见表6，表中部分数据已做脱敏处理。

对表6所列的关联规则逐一分析，可得到该局机车的事故故障发生情况及质量安全状态如下。

（1）所有事务中，项目“机破发生1次”与“H1机务段”共同发生286次，占所有事务的比例为0.91%，“机破发生1次”对于“H1机务段”的置信度高达90.22%，提升度为3.12，表示每月发生1次机破的机车有90.22%的比例集中在H1机务段，这条关联规则说明H1机务段的机车每月较容易发生1次机破问题，因此该段在机车质量安全

表5 重点关注项目的最小支持度索引表

序号	重点关注项目	最小支持度计数值	序号	重点关注项目	最小支持度计数值
1	机破高发	10	15	春鉴等级合格	170
2	运用故障高发	20	16	D级机车	200
3	机破发生2次	20	17	零公里质量鉴定问题很多	220
4	性能试验问题过多	30	18	机破发生1次	250
5	C6	30	19	临修高发	350
6	大修	30	20	行车安全装备问题过多	450
7	C5	30	21	碎修超多	600
8	春鉴等级不合格	40	22	必换件问题发生	900
9	运用故障较多	50	23	碎修过多	950
10	零公里质量鉴定问题过多	70	24	行车安全装备问题较多	1 000
11	五项专检问题过多	80	25	五项专检问题很多	1 200
12	性能试验问题较多	110	26	C级机车	1 500
13	C4	150	27	春鉴等级良好	1 500
14	中修	150	28	B级机车	5 000

表6 挖掘形成的6条代表性关联规则

序号	关联规则	关联规则在事务集合中出现的频次/次	支持度/%	置信度/%	提升度
1	机破发生1次⇒H1机务段	286	0.91	90.22	3.12
2	运用故障高发⇒行车安全装备问题较多	31	0.10	83.78	21.27
3	D级机车⇒碎修超多	247	0.79	65.52	22.83
4	性能试验问题较多⇒H3机务段,C级机车	85	0.27	54.49	18.68
5	零公里质量鉴定问题过多⇒H2机务段,B级机车	93	0.30	60.78	8.51
6	五项专检问题过多⇒H3机务段,B级机车,碎修较多	103	0.33	60.23	14.12

管理中，应对机车机破问题的防范加以关注。

（2）项目“运用故障高发”与“行车安全装备问题较多”共同发生31次，其置信度为83.78%，提升度达到了21.27，表明机车运用故障的高发往往伴随着较多的行车安全装备问题，二者间的关联性很强。机车的行车安全装备包含机车信号、列车运行监控记录装置、机车综合无线通信设备、机车车载安全防护系统等多种设备，虽然这条关联规则的支持度仅有0.10%，在所有事务中同时出现的占比不高，但考虑到运用故障和行车安全装备问题对于机车的运行安全有着较为严重的影响，因此机车日常养护维修作业中，应加强对于行车安全装备的质量把控，减少运用故障的高发对机车行车安全的影响。

（3）项目“D级机车”与“碎修超多”共同发生247次，“D级机车”对于“碎修超多”的置信度为65.52%，二者的提升度高达22.83，表明该局总体质量最差的D级机车与机车碎修问题有很强的关联性，该局65.52%的D级机车经常表现出超多的碎修问题，因此在对全局质量状态最差的D级机车开展专项整治工作时，可从运用质量维度下的碎修问题着手，有针对性地安排相应的检修修程，减少机车故障数量，提升全局机车质量。

（4）项目“性能试验问题较多”“H3机务段”和“C级机车”共同发生85次，“性能试验问题较多”对于“H3机务段”和“C级机车”的置信度为54.49%，提升度为18.68。表明全局发生较多性能试验问题的机车，有54.49%的比例集中于H3机务段的C级机车，且表现出很强的关联性。因此，该段在对总体质量较差的C级机车进行检修维护作业时，可从机车检修工作中的性能试验入手，加强性能试验所涉及的制动机、高低压、负载等机车设备的检查维修力度，防范机车设备故障的发生，促进C级机车的质量提升。

（5）项目“零公里质量鉴定问题过多”“H2机务段”和“B级机车”共同发生93次，“零公里质量鉴定问题过多”对于“H2机务段”和“B级机车”的置信度达到60.78%，提升度为8.51，表明H2机务段的B级机车更容易发生零公里质量鉴定问题，因此该段在机车检修作业中，可针对零公里

质量鉴定中发现的机车质量问题，有的放矢地加强相应机车设备的维护保养，减轻该问题对B级机车的影响。

（6）项目“五项专检问题过多”“H3机务段”“B级级车”和“碎修较多”共同发生103次，“五项专检问题过多”对于“H3机务段”“B级级车”和“碎修较多”的置信度为60.23%，提升度为14.12，表明H3机务段的B级机车中，“五项专检问题过多”与“碎修较多”这2项机车质量问题之间存在较强的关联性。机车五项专检涵盖了走行部、车顶高压设备、DC 600V直流供电、制动机、防火等多种设备，因此该段应针对性地开展跨运输生产环节的专项整治工作，减少五项专检问题和碎修问题的发生。

6　结　语

本文立足铁路机务专业的机车事故故障关联分析需要，采用Eclat算法的垂直数据挖掘思想，通过改善其无法满足多最小支持度关联规则挖掘的缺陷，提出了MsEclat这一改进算法，详细阐释了MsEclat算法的数据挖掘思路，并给出示例。进一步地，为更好实现大数据场景下的关联规则挖掘，利用布尔矩阵和并行计算编程模型MapReduce对MsEclat算法加以优化，形成优化的MsEclat算法，设计了相应的频繁项集挖掘步骤。选取数据规模和数据稠密程度各不相同的3个数据集，将MsEclat算法与MsApriori算法、MS-Growth算法这2种水平挖掘算法以及优化的MsEclat算法分别进行比较，证明MsEclat算法及其优化算法在多最小支持度关联规则挖掘的执行效率上均有着极好的表现，特别是优化的MsEclat算法，处理大规模数据时的执行效率得到进一步提高。将优化的MsEclat算法应用到某铁路局的机车事故故障关联分析这一具体的大数据关联规则挖掘场景，得到了与机车事故故障这类重点数据相关的多条关联规则，如运用故障的高发有83.78%的可能性会伴随着较多的行车安全装备问题等等，并分别做出了相应分析，证实该算法对科学、高效、精准地开展铁路设备质量安全状态分析具有良好的技术支撑作用。

参　考　文　献

[1] 史天运，刘军，李平，等.铁路大数据平台总体方案及关键技术研究[J].铁路计算机应用，2016，25(9)：1-6.
（SHI Tianyun，LIU Jun，LI Ping，et al. Overall Scheme and Key Technologies of Big Data Platform for China Railway [J]. Railway Computer Application，2016，25 (9)：1-6. in Chinese）

[2] 谯兵，胡斌.基于EHM理念的铁路机务设备大数据健康管理系统的设计与实现[J].铁路计算机应用，2019，28(12)：35-39.
（QIAO Bing，HU Bin. Big Data Health Management System of Railway Locomotive Equipment Based on EHM Concept [J]. Railway Computer Application，2019，28 (12)：35-39. in Chinese）

[3] 崔妍，包志强.关联规则挖掘综述[J].计算机应用研究，2016，33(2)：330-334.
（CUI Yan，BAO Zhiqiang. Survey of Association Rule Mining [J]. Application Research of Computers，2016，33 (2)：330-334. in Chinese）

[4] 晏杰，亓文娟，郭磊，等.基于多最小支持度的关联规则挖掘[J].计算机系统应用，2014，23(3)：237-239，219.
（YAN Jie，QI Wenjuan，GUO Lei，et al. Based on Multiple Minimum Supports of Association Rules in Data Mining [J]. Computer Systems & Applications，2014，23 (3)：237-239，219. in Chinese）

[5] 常浩，陈莉.多最小支持度关联规则挖掘研究[J].微计算机信息，2010，26(24)：143-144，5.
（CHANG Hao，CHEN Li. Research on the Multiple Minimum Supports Association Rules Mining [J]. Microcomputer Information，2010，26 (24)：143-144，5. in Chinese）

[6] 王海波，张永田，吴升.基于数据立方体的多最小支持度关联规则在犯罪分析中的应用[J].测绘科学技术学报，2016，33(4)：405-409.
（WANG Haibo，ZHANG Yongtian，WU Sheng. Application of Association Rules with Multiple Minimum Supports Based on the Data Cube in Crime Analysis [J]. Journal of Geomatics Science and Technology，2016，33 (4)：405-409. in Chinese）

[7] HAN J W, PEI J, YIN Y W, et al. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach [J]. Data Mining and Knowledge Discovery, 2004, 8 (1): 53-87.

[8] HU Y H, CHEN Y L. Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism [J]. Decision Support Systems, 2006, 42 (1): 1-24.

[9] 张慧哲,王坚.多重最小支持度频繁项集挖掘算法研究[J].计算机应用,2007,27(9):2290-2293.
(ZHANG Huizhe, WANG Jian. Research of Multiple Minimum Supports Frequent Itemsets Mining [J]. Journal of Computer Applications, 2007, 27 (9): 2290-2293. in Chinese)

[10] 魏恩超.基于紧凑模式树和多最小支持度的频繁模式挖掘算法研究[D].西安:西安理工大学,2019.
(WEI Enchao. Research Frequent Pattern Mining Algorithm Based on Compact Pattern Tree and Multiple Minimum Support [D]. Xi'an: Xi'an University of Technology, 2019. in Chinese)

[11] 向春梅,陈超.基于MapReduce的改进Eclat算法[J].成都信息工程大学学报,2019,34(4):369-374.
(XIANG Chunmei, CHEN Chao. Improved Eclat Algorithm Based on MapReduce [J]. Journal of Chengdu University of Information Technology, 2019, 34 (4): 369-374. in Chinese)

[12] CHOUBEY A, PATEL R, RANA J L. A Survey of Efficient Algorithms and New Approach for Fast Discovery of Frequent Itemset for Association Rule Mining (DFIARM) [J]. International Journal of Soft Computing and Engineering, 2011, 2 (1): 62-67.

[13] ZAKI M J, GOUDA K. Fast Vertical Mining Using Diffsets [C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, D. C.. New York: Association for Computing Machinery Press, 2003: 326-335.

[14] HAN J W, KAMBER M. Data Mining: Concepts and Techniques [M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2011.

[15] 梁杨,钱晓东.多最小支持度关联规则改进算法[J].西南大学学报:自然科学版,2019,41(7):131-141.
(LIANG Yang, QIAN Xiaodong. An Improved Algorithm for Association Rules with Multiple Minimum Supports [J]. Journal of Southwest University: Natural Science Edition, 2019, 41 (7): 131-141. in Chinese)

[16] 张玉芳,熊忠阳,耿晓斐,等.Eclat算法的分析及改进[J].计算机工程,2010,36(23):28-30.
(ZHANG Yufang, XIONG Zhongyang, GENG Xiaofei, et al. Analysis and Improvement of Eclat Algorithm [J]. Computer Engineering, 2010, 36 (23): 28-30. in Chinese)

[17] 王燕.基于等价关系的关联规则挖掘算法研究[J].计算机工程与应用,2006,42(8):187-189.
(WANG Yan. Algorithm Research for Mining Association Rule Based on Equivalence Relation [J]. Computer Engineering and Applications, 2006, 42 (8): 187-189. in Chinese)

[18] 王红梅,胡明,赵守峰.基于垂直格式的频繁项集挖掘分段算法[J].吉林大学学报:理学版,2016,54(3):553-560.
(WANG Hongmei, HU Ming, ZHAO Shoufeng. Frequent Itemsets Mining Segmentation Algorithm Based on Vertical Format [J]. Journal of Jilin University: Science Edition, 2016, 54 (3): 553-560. in Chinese)

[19] 熊忠阳,陈培恩,张玉芳.基于散列布尔矩阵的关联规则Eclat改进算法[J].计算机应用研究,2010,27(4):1323-1325.
(XIONG Zhongyang, CHEN Peien, ZHANG Yufang. Improvement of Eclat Algorithm for Association Rules Based on Hash Boolean Matrix [J]. Application Research of Computers, 2010, 27 (4): 1323-1325. in Chinese)

[20] LIANG S Y. Research on the Method and Application of MapReduce in Mobile Track Big Data Mining [J]. Recent Advances in Electrical & Electronic Engineering: Formerly Recent Patents on Electrical & Electronic Engineering, 2021, 14 (1): 20-28.

[21] KHEZR S N, NAVIMIPOUR N J. MapReduce and Its Applications, Challenges, and Architecture: A Comprehensive Review and Directions for Future Research [J]. Journal of Grid Computing, 2017, 15 (3): 295-321.

Association Rule Mining for Railway Locomotive Accident and Fault Based on Optimized MsEclat Algorithm

LI Xin[1], SHI Tianyun[2], CHANG Bao[3], MA Xiaoning[3], LIU Jun[3]

(1. Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China;
2. Department of Science, Technology and Information Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China;
3. Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: In order to mine the association rules related to locomotive accidents and faults from railway locomotive big data, an optimized MsEclat algorithm is proposed. Firstly, an improved Eclat algorithm, MsEclat algorithm, is proposed. By constructing the minimum support index table, the data set is reconstructed according to the support value of each item as ordering, and the frequent item sets for different items are obtained according to the vertical mining idea, to solve the defect that the Eclat algorithm cannot mine association rules in case of multiple minimum support. Secondly, the optimized MsEclat algorithm is further improved. Based on the integration of Boolean matrix and parallel computing programming model MapReduce, frequent item set mining steps are designed to improve the execution efficiency of algorithm in big data analysis scenarios. The comparison of different algorithms shows that MsEclat algorithm and its optimization algorithm have great advantages in computing efficiency for mining association rules with multiple minimum supports. Finally, taking the big data of locomotive operation and maintenance of a railway bureau as an example, the optimized MsEclat algorithm is used to mine the association rules of locomotive accidents and faults. The results show that the optimized MsEclat algorithm takes 3.945 034 s in the case of 6 distributed nodes, and 156 frequent item sets are mined. For example, in the use of locomotives with high frequency operation faults, 83.78% of the probability will simultaneously appear more driving safety equipment problems. After forming these association rules, they can be used to analyze the occurrence of accidents and faults as Well as the quality and safety state of locomotives in the railway bureau.

Key words: Locomotive accident and fault; Association rule; Big data analysis; Data mining technology; MsEclat algorithm; Multiple minimum supports

（责任编辑　刘卫华）