

文章编号 1000-5013(2004)02-0118-03

关联规则在倒班运行质量分析中的应用

林 峰

(华侨大学数学系, 福建 泉州 362011)

摘要 研究一个倒班运行质量分析问题, 并应用关联规则发现的方法, 产生运行班组与运行质量评价指标之间的关联规则, 用于评判班组的运行质量. 针对问题的特殊性, 对关联规则发现的 Apriori 算法作了改进, 并给出在火电厂锅炉运行质量分析中应用的例子.

关键词 数据挖掘, 关联规则, 倒班运行

中图分类号 TP 311.132; TK 227

文献标识码 A

在倒班运行中, 同一天中是由多个班组轮流操作设备的. 不同班组人员在操作设备时, 其技术水平、责任心等各种因素都可能各不相同. 这就可能存在着运行质量好坏的问题, 因此人们希望利用日运行记录数据来分析评判班组的运行质量. 但是, 困难在于日运行记录数据所反映的是当天参与运行的多个班组的综合结果. 如何运用日运行记录数据, 分析或评判单个班组的运行质量的好坏, 这是我们多年前曾经考虑的一个问题, 同时也曾尝试过许多方法而难以得到结果. 关联规则发现是数据挖掘中研究的重要课题之一^[1]. 随着关联规则发现方法得到越来越广泛的应用^[2, 3], 关联规则的概念和意义也随之得到越来越多的人们的理解和接受. 这为研究和解决实际问题奠定了基础. 本文应用关联规则发现方法, 找出运行班组与运行质量评价指标之间的关联规则, 并用于评判班组运行的质量好坏. 从在实际例子中的应用来看, 其结果令人满意.

1 关联规则的概念

设 $I = \{I_1, I_2, \dots, I_m\}$ 是项的集合. 事务 T 是 I 的子集, 不同的事务的集合构成数据库 D . 关联规则的形式为 $X \Rightarrow Y$, 其中 $X \subset I$, $Y \subset I$ 并且 $X \cap Y = \Phi$. 这里 X, Y 称为项集. 任意一个项集 X 的支持度, 可以定义为

$$\text{support}(X) = |\{T \in D \mid X \subseteq T\}| / |D|,$$

其中 $|\cdot|$ 表示集合中元素的个数.

任意一条规则 $X \Rightarrow Y$ 的支持度和置信度分别定义为

$$\text{support}(X \Rightarrow Y) = \text{support}(X \cup Y),$$

$$\text{confidence}(X \Rightarrow Y) = \text{support}(X \Rightarrow Y) / \text{support}(X).$$

关联规则发现, 就是生成支持度与置信度都分别不小于事先给定的最小值的所有关联规则.

2 倒班运行质量分析问题

对于 n 班 k 倒制的倒班运行, 共有 n 个班组. 每天上班的仅为其中的 k 个班组, 记 n 个班组为 x_1, x_2, \dots, x_n . 运行质量评价指标可分为若干档次, 设共有 r 档, 记为 y_1, y_2, \dots, y_r . 每天由 k 个班组 $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ ($1 \leq i_1, i_2, \dots, i_k \leq n$) 和质量评价指标的某个档次 y_j ($1 \leq j \leq r$), 作为项构成一个事务 T , 即 $T = \{x_{i_1}, x_{i_2}, \dots, x_{i_k}, y_j\}$.

收稿日期 2003-07-27

作者简介 林 峰(1962-), 男, 高级讲师, 主要从事函数论和数据挖掘的研究. E-mail: lfeng@hqu.edu.cn

数据库 D 由不同的事务 T 的集合构成. 解决倒班运行质量分析问题, 就是首先生成支持度与置信度, 都分别不小于事先给定的最小值的如下形式的关联规则. 即

$$X \Rightarrow Y, \quad X \subset \{x_1, x_2, \dots, x_n\}; \quad Y \in \{y_1, y_2, \dots, y_r\}.$$

(1)

然后, 根据生成的这些关联规则, 分析和评判单个班组或班组的组合的运行质量的好坏.

3 算法的研究

对一般的关联规则发现问题, 通常采用著名的 Apriori 算法. 如果直接应用于上述倒班运行质量分析问题, 得到的将是如下形式的关联规则. 即

$$X \Rightarrow Y, \quad X, Y \subset \{x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_r\},$$

(2)

其中包含了许多无用的关联规则, 且在计算过程中, 时间和空间也有极大的浪费. 因此, 有必要针对问题的特殊性对算法进行改进.

如果在关联规则发现问题中, 项的集合 I 具有特殊的构成 $I = \{x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_r\}$. 事务数据库由具有的形式, 由 $T = \{x_{i1}, x_{i2}, \dots, x_{ik}, y_{ij}\}$ 的事务构成. 要发现的关联规则形式为

$$X \Rightarrow Y, \quad X \subset \{x_1, x_2, \dots, x_n\}; \quad Y \in \{y_1, y_2, \dots, y_r\}.$$

(3)

那么这将是具有特殊性的关联规则发现问题. 它不仅可以应用于本文的倒班运行质量分析问题, 也必定还有其他的应用. 因此, 研究其算法是有意义的.

通过对 Apriori 算法的改进, 得到这一类特殊的关联规则发现问题的算法. (1) 输入. 事务数据库 D ; 最小支持度阈值 $\min\text{-sup}$; 最小置信度阈值 $\min\text{-conf}$. (2) 输出. 所产生的关联规则集 R_k .

下面介绍算法的具体步骤. (1) 计算候选 1-项集 C_1 有, $C_1 = \{x_1, x_2, \dots, x_n\}$. (2) 计算频繁 1-项集 L_1 及产生规则集 R_1 . 扫描 C_1 , 对于 x_i , 如果 $\text{support}(x_i \Rightarrow y_j) < \min\text{-sup}$ 对每个 y_j 成立, 则 x_i 称为非频繁项集. 将 x_i 从 C_1 中删除, 否则, 如果 $\text{confidence}(x_i \Rightarrow y_j) \geq \min\text{-conf}$ 对某个 y_j 成立. 那么, 即将规则 $x_i \Rightarrow y_j$ 加入规则集 R_1 中. C_1 删除所有非频繁项集后构成 L_1 . (3) k 赋值为 2. (4) 计算候选 k -项集 C_k . 与 Apriori 算法相同, 由 L_{k-1} 经过连接步产生所有 k -项集, 构成 C_k . 再经过剪枝步对 C_k 压缩, 即检查每一个 k -项集的 $(k-1)$ -子集是否都属于 L_{k-1} . 若有一个 $(k-1)$ -子集不属于 L_{k-1} , 则将该 k -项集从 C_k 中删除. (5) 计算频繁 k -项集 L_k 及产生规则集 R_k . 扫描 C_k , 对 k -项集 x , 如果 $\text{support}(x \Rightarrow y_j) < \min\text{-sup}$ 对每个 y_j 成立, 则 x 称为非频繁项集. 将 x 从 C_k 中删除, 否则, 如果 $\text{confidence}(x \Rightarrow y_j) \geq \min\text{-conf}$ 对某个 y_j 成立, 则将规则 $x \Rightarrow y_j$ 加入规则集 R_k 中. C_k 删除所有非频繁项集后构成 L_k . (6) $k = k + 1$. (7) 循环调用步骤 (4)~(6), 直到 C_k 为空. (8) 结束.

4 应用举例

本例对福建省某火力发电厂 100 MW 发电机组的锅炉运行, 进行班组运行质量分析^[4]. 该运行采用 5 班 4 倒制. 选取厂用电损耗比例作为评价指标, 分别有球磨耗 Q_M 、排粉耗 P_F 、送风耗 S_F 、引风耗 Y_F . 利用某个连续阶段满负荷发电的 51 d 的日数据, 原始数据示意如表 1(表中 Q 表示发电量).

表 1 原始数据示意表

序号	运行班组	$Q/\text{MW}\cdot\text{h}$	$Q_M/\text{MW}\cdot\text{h}$	$P_F/\text{MW}\cdot\text{h}$	$S_F/\text{MW}\cdot\text{h}$	$Y_F/\text{MW}\cdot\text{h}$
1	1, 2, 3, 4	2 346	30. 720	13. 440	24. 360	21. 240
2	2, 3, 4, 5	2 381	29. 280	13. 080	24. 480	21. 480
...
51	1, 2, 4, 5	2 379	31. 920	13. 800	23. 400	21. 000

4.1 发现关联规则

对于表 1 中的每一行, 可按以下公式计算厂用电损耗比例. 即

$$\text{厂用电损耗比例} = \frac{\text{球磨耗} + \text{排粉耗} + \text{送风耗} + \text{引风耗}}{\text{发电量}}.$$

然后求出厂用电损耗比例的平均值. 对厂用电损耗比例小于平均值的, 当天 4 个班组的运行质量评价指

标归档为“好”,用记号 y_1 表示.对厂用电损耗比例大于等于平均值的,当天 4 个班组的运行质量评价指标归档为“差”,用记号 y_2 表示.这样可以构造出事务数据库 D ,示意表如表 2 所示.给定支持度最小值 $\text{minsup}=10\%$,置信度最小值 $\text{minconf}=70\%$.在事务数据库 D 中进行关联规则发现,产生的关联规则如表 3 所示.

表 2 D 中事务示意表		表 3 产生的关联规则			
序号	事 务	序号	关联规则	规则支持度/(%)	规则置信度/(%)
1	$\{x_1, x_2, x_3, x_4, y_2\}$	1	$\{x_1, x_5\} \Rightarrow \{y_1\}$	39	71
2	$\{x_2, x_3, x_4, x_5, y_1\}$	2	$\{x_1, x_3, x_5\} \Rightarrow \{y_1\}$	27	74
...	...	3	$\{x_1, x_2, x_5\} \Rightarrow \{y_1\}$	25	81
51	$\{x_1, x_2, x_4, x_5, y_2\}$	4	$\{x_1, x_2, x_3, x_5\} \Rightarrow \{y_1\}$	14	100

4.2 分析评判

从产生出的 4 条关联规则来看,第 1 班和第 5 班都分别有 4 条关联规则与运行质量指标“好”关联.这说明第 1 班和第 5 班运行质量最好.第 2 班和第 3 班都分别有 2 条关联规则与运行质量指标“好”关联.这说明第 2 班和第 3 班运行质量较好.第 4 班没有与运行质量指标“好”关联的规则,说明第 4 班运行质量差.没有产生与运行质量指标“差”关联的规则,甚至第 4 班也没有.这是因为每天参与运行的有 4 个班,一个“差”的班加上其他“好”或“较好”的班后,综合效果并不“差”.

5 结束语

一般认为,关联规则发现方法起源于对商务数据库的分析.例如,购物篮分析、客户分析等.目前关联规则发现方法的应用虽然逐渐广泛,但象本文把关联规则应用于倒班运行质量分析中,乃是一个大胆的尝试.我们在实例中,成功发现了 4 条关联规则让人感到鼓舞.应用所发现的关联规则,对班组运行质量进行评价的结果也令人信服.然而,应用所发现的关联规则对班组运行质量进行评价,在不同的问题中还可能出现不同的情况.如何结合实际情况进行分析评判,还需要到更多的实践中加以总结.

参 考 文 献

1 Han J, Kamber M 著.数据挖掘概念与技术[M]. 范 明等译.北京:机械工业出版社,2001. 1~ 200
2 丁 夷.关联规则挖掘在电信市场研究中的应用[J]. 西安邮电学院学报,2000, 5(3): 39~ 41
3 孙建平,梅晓勇,史忠植等.关联规则在高校智能排课系统中的应用[J]. 计算机应用,2002, 22(5): 37~ 39
4 林 峰.用数学方法确定锅炉指标最佳值的探讨[J]. 福建电力与电工,1997, 17(2): 45~ 46

Application of Association Rule to the Operation of
Work in Shifts for Its Quality Analysis

Lin Feng

(Dept. of Math. Huaqiao Univ., 362011, Quanzhou, China)

Abstract A study is made on the quality analysis given to the operation of work in shifts. For the use of judging the operation of work in shifts, the association rule between working teams and index for evaluating quality of operation is engendered by applying the method for discovering association rule. Aiming at the particularity of this problem, the author improves Apriori algorithm by which association rule is discovered; and gives an example of application to the operation of the boiler in a thermal power plant for its quality analysis.

Keywords data mining, association rule, operation of work in shifts