关联规则挖掘在电厂设备故障监测中应用

梁志瑞,陈 鹏,苏海锋
(华北电力大学 电气工程学院,河北 保定 071003)

摘要:关联规则挖掘是数据挖掘的重要分支,其通过描述数据库中不同数据属性之间所存在的潜在关系规则,找出满足给定支持度阈值和置信度阈值多个域之间的依赖关系。随着电厂设备运行期间各种故障的发生,各状态监测点参数也会发生相应变化,利用关联规则挖掘算法,找出故障发生时故障现象与故障类别之间的关联关系,更好地对设备进行故障监测与诊断。阐述了关联规则挖掘的主要概念,对挖掘时最常用的 Apriori 算法进行探讨,并以汽轮机凝汽器的一种典型故障为例说明了算法的执行情况,对挖掘结果进行了解释。结果验证了所用方法的可行性与正确性。

关键词:关联规则;Apriori 算法;故障监测
中图分类号:TM 621;TP 311 **文献标识码**:A **文章编号**:1006-6047(2006)06-0017-03

随着计算机应用普及,电厂设备管理的自动化水平得以不断提高,各种设备管理软件已在现场被广泛使用。设备状态监测模块作为设备管理系统的重要组成部分,随着时间推移将会积累大量历史数据,这其中包含了许多故障时各监测点的测量参数。如何对这些记录加以分析利用,找出其中隐藏的规律,对设备的故障监测有着重要意义。

在数据挖掘的知识模式中,关联规则挖掘是比较重要的一种。关联规则挖掘侧重于寻找给定数据集中不同数据属性之间的联系。通过描述数据库中数据项之间所存在的潜在关系的规则,找出满足给定支持度阈值和置信度阈值的多个域之间的依赖关系。设备发生故障的同时,一些故障现象会伴随着发生。利用故障时各个监测点的数据,运用关联规则挖掘算法,可找出在某个故障发生时,故障现象与故障类别之间的关系,以这些故障现象为依据可对事故发生进行及时预警,而且故障时各关联监测点的参数值也可为故障诊断提供依据。

1 关联规则概念描述[1-2]

设要进行关联规则挖掘的数据集记为 D(D 为事务数据库),$D=\{t_1,t_2,\cdots,t_k,\cdots,t_n\}$,其中的任意元素 $t_k=\{i_1,i_2,\cdots,i_j,\cdots,i_p\}$($k=1,2,\cdots,n$)都对应 1 个事务,$t_k$ 中的元素 i_j($j=1,2,\cdots,p$)称为项目。设 $I=\{i_1,i_2,\cdots,i_m\}$ 是所有项目的集合,I 的任意子集 X 称为 D 中的项目集,若 $|X|=k$ 则称集合 X 为 k-项目集。设 t_k 和 X 分别为 D 中的事务和项目集,如果 $X\subseteq t_k$,则称事务包含项目集。数据集 D 中包含项目集 X 的事务数称为项目集 X 的支持数,记为 σ_x,项目集 X 的支持度记作 support(X)。

$$\text{support}(X)=\frac{\sigma_x}{|D|}\times 100\%$$

式中 $|D|$ 为数据集 D 的事务数。

若 support(X)不小于用户指定的最小支持度(记为 minsupport),则称 X 为频繁项目集(或大项目集),否则称 X 为非频繁项目集(或小项目集)。

若 X、Y 为项目集,且 $X\cap Y=\varnothing$,则蕴涵式 $X\Rightarrow Y$ 称为关联规则,项目集($X\cup Y$)的支持度称为关联规则 $X\Rightarrow Y$ 的支持度,是 D 中事务包含($X\cup Y$)的百分比,即概率 $P(X\cup Y)$,记为 support($X\Rightarrow Y$)。

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

关联规则 $X\Rightarrow Y$ 的置信度是 D 中事务包含 X 的事务的同时也包含 Y 的百分比,即条件概率 $P(Y/X)$,记为 confidence($X\Rightarrow Y$)。

$$\text{confidence}(X\Rightarrow Y)=\frac{\text{support}(X\cup Y)}{\text{support}(X)}\times 100\%=P(Y/X)$$

通常用户应根据实际需要来指定最小支持度和最小置信度,前者表示的是数据项集在统计意义上的最低主要性,后者表示的是规则的最低可靠性,称支持度和置信度都不小于其相应阈值的规则称为强关联规则。关联规则挖掘的主要任务就是从数据库中发现有实际意义的强关联规则。

根据上述定义,关联规则挖掘问题可以分解为以下 2 个子问题:

a. 找出所有频繁项目集,根据定义,这些项目集出现的频率应不小于最小支持度;

b. 由频繁项目集产生强关联规则,根据定义,这些规则必须满足最小支持度和最小置信度。

其中第 1 个问题是关键,下面介绍一种具有代表性的寻找频繁项目集的算法——Apriori 算法。

2 Apriori 算法描述

基本算法 Apriori 提出了挖掘关联规则的基于 2 阶段频繁集的方法,算法描述如下[3-5]。

收稿日期:2005-09-09;修回日期:2005-12-19

a. 输入：事物数据库 D；最小支持度阈值 min-support。

b. 输出：D 中所有强关联规则的集合 R。

c. 算法：

F_1=find_frequent_1-itemset(D);

//扫描数据集 D，找出所有频繁 1 项集的集合 F_1

$F=\emptyset$; //清空输出集合

for(k=2; $F_{k-1}\neq\emptyset$;k++){

C_k= apriori_gen(F_{k-1},minsupport);

for each transaction $t\in D$ {

//扫描 D 中的每条事务进行候选计数

C_t= subset(C_k,t); //取得由事务 t 产生的候选集合

for each candidate $c\in C_t$

c.count ++;} //对候选集合中的每个候选计数加 1

$F_k=\{c\in C_k \mid \frac{c.\text{count}}{|D|}\geqslant \text{minsupport}\}$;}

return $F=\cup_k F_k$; //得到频繁项目集的集合

R=generate_rule (F); //由频繁项目集产生关联规则

R return(R);

procedure apriori_gen(F_{k-1}:frequent(k-1)-itemsets;

minsupport:minimum support threshold)

$C_k=\emptyset$;

for each itemset $f_1\in F_{k-1}$ //每个 F_{k-1} 中的项目集 f_1

for each itemset $f_2\in F_{k-1}$ //每个 F_{k-1} 中的项目集 f_2

if(($f_1[1]=f_2[1]$)∧($f_1[2]=f_2[2]$)∧…∧($f_1[k-2]=f_2[k-2]$)∧($f_1[k-1]<f_2[k-1]$))

then{

$c=\{f_1[1],f_1[2],\cdots,f_1[k-1],f_2[k-1]\}$;

//连接步，产生候选 k- 项目集

if has_infrequent_subset(c,F_{k-1}) then

delete c; //剪枝步

else add c to C_k;}

return C_k;

procedure has_infrequent_subset(c: candidate k-itemset;F_{k-1}:frequent(k-1)-itemset)

for each(k-1)- subset s of c //c 中的每一个 k-1 子项目集

if $s\notin F_{k-1}$ then

return TRUE;

else return FALSE;

需要注意：在剪枝步，应用了 Apriori 算法隐含的一个重要性质，即最大项目集的子集必为最大项目集。

3 实际应用

以汽轮机凝汽器的一种典型故障为例[6-7]，说明如何运用关联规则挖掘的 Apriori 算法找出故障现象与故障类别之间的关联关系。

设某火力发电厂 200 MW 单元发电机组的汽轮机凝汽器发生后轴封供汽中断故障时出现过 5 种故障现象，分别是真空急剧下降、转子出现负胀差、凝汽器端差增加、凝结水过冷度增加和循环水温升减小（这里的增加或减小是指用参数当前负荷实际运行值与应达目标值比较并考虑不同的偏差阈值），分别记为 A~E。收集凝汽器每次发生后轴封供汽中断故障时的故障记录，如表 1 所示。

表 1 故障现象记录表

Tab.1 List of recorded fault phenomena

记录编号	故障现象	记录编号	故障现象
1	A,B,C	6	A,C,D
2	A,B,D	7	A,B,D
3	B,C	8	A,B
4	A,B,D,E	9	A,B,C,D
5	A,B		

表 1 所示即为算法中的事物数据库 D，在此给定最小支持度为 0.3，即最小支持度计数为 3，则按照 Apriori 算法进行关联规则挖掘的步骤如图 1 所示[8-9]。

根据图 1 结果，可得到如下的关联规则：

a. 发生故障时，故障现象 A 出现的同时故障现象 B 出现的置信度为 7/8=87.5 %；

b. 发生故障时，故障现象 A 出现的同时故障现象 C 出现的置信度为 3/8=37.5 %；

c. 发生故障时，故障现象 A 出现的同时故障现象 D 出现的置信度为 5/8=62.5 %；

d. 发生故障时，故障现象 B 出现的同时故障现象 C 出现的置信度为 3/8=37.5 %；

e. 发生故障时，故障现象 B 出现的同时故障现象 D 出现置信度为 4/8=50 %；

f. 发生故障时，故障现象 A 出现的同时故障现象 B、D 同时出现的置信度为 4/8=50 %；

g. 发生故障时，故障现象 B 出现的同时故障现象 A、D 同时出现的置信度为 4/8=50 %；

h. 发生故障时，故障现象 D 出现的同时故障现象 A、B 同时出现的置信度为 4/5=80 %；

i. 发生故障时，故障现象 A、B 出现的同时故障现象 D 出现的置信度为 4/7=57 %；

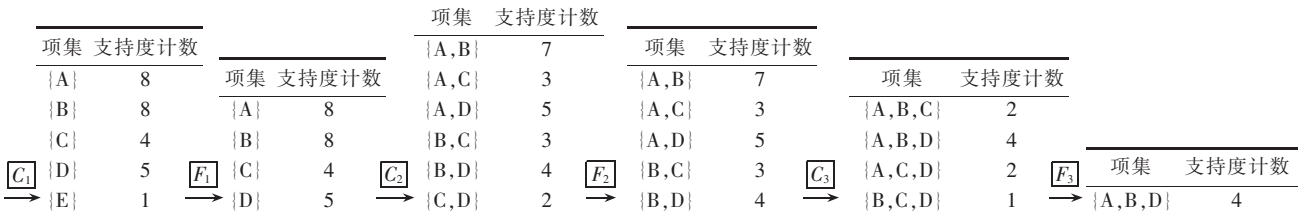

图 1 频繁项目集的产生过程

Fig.1 Generation procedure of large item sets

j. 发生故障时,故障现象 A、D 出现的同时故障现象 B 出现的置信度为 4/5=80 %;

k. 发生故障时,故障现象 B、D 出现的同时故障现象 A 出现的置信度为 4/4=100 %。

若指定最小置信度为 85 %,则只剩下规则 **a**、**k**。规则 **a** 表达如下信息:即在汽轮机凝汽器发生后轴封供汽中断故障时,在发生真空急剧下降的情况下常发生转子出现负胀差的情况。同理,规则 **k** 表达信息如下:即在同类型故障时,同时发生转子出现负胀差、凝结水过冷度增加的情况下,一定会发生真空急剧下降的情况。由此可根据这些监测点的参数变化,对该类故障进行预警;还可将故障时这些测点的即时参数值作为故障诊断专家系统的模型数据[10]。

4　结语

电厂设备管理系统包含了多个模块,设备状态监测只是其中的一个组成部分,应用关联规则挖掘算法可从中挖掘故障现象与故障类别之间的规律,为今后的故障监测提供依据。本文仅就关联规则挖掘的一种经典算法在电厂设备故障监测中的应用作了简单探讨,如何将其他的数据挖掘技术引用到电厂设备管理系统中,找出有可能隐含的规律,并与实际情况相结合,进一步指导实际工作,还需要大量工作。

参考文献:

[1] 邵峰晶,于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社,2003.

[2] 刘同名. 数据挖掘技术及其应用[M]. 北京:国防工业出版社,2001.

[3] 曾海颖. 客户关系管理中的数据挖掘[D]. 南京:南京航空航天大学,2003.
ZENG Hai-ying. Data mining in customer relationship management [D]. Nanjing:Nanjing University of Aeronautics and Astronautics, 2003.

[4] 蒋秀英. 关联规则在课堂教学评价中的应用[J]. 山东师范大学学报,2003,18(3):25-28.
JIANG Xiu-ying. The implementation of association rules in classroom teaching evaluation[J]. Journal of Shandong Normal University,2003,18(3):25-28.

[5] KANTARDZIC M. 数据挖掘——概念、模型、方法和算法[M]. 闪四清,陈茵,程雁,等,译. 北京:清华大学出版社,2003.

[6] 马良玉,王兵树,佟振声,等. 对分式凝汽器故障诊断的模糊模式识别及神经网络方法[J]. 中国电机工程学报,2001,21(8):68-73.
MA Liang-yu,WANG Bing-shu,TONG Zhen-sheng,et al. Fuzzy pattern recognition and artificial neural network used for fault diagnosis of the double channel condenser[J]. Proceedings of the CSEE,2001,21(8):68-73.

[7] 王培红,朱玉娜,贾俊颖,等. 模糊模式识别在凝汽器故障诊断中的应用[J]. 中国电机工程学报,1999,19(10):46-49.
WANG Pei-hong,ZHU Yu-na,JIA Jun-ying,et al. Application of fuzzy pattern recognition[J]. Proceedings of the CSEE,1999,19 (10):46-49.

[8] 王选文,丁夷,范九伦. 关联规则挖掘在人事系统中的应用[J]. 西安邮电学院学报,2001,6(1):183-186.
WANG Xuan-wen,DING Yi,FAN Jiu-lun. The implementation of association rules in people management system[J]. Journal of Xi'an Institute of Posts and Telecommunications,2001,6(1):183-186.

[9] 陆丽娜,陈亚萍,魏恒义,等. 挖掘关联规则掘中 Apriori 算法的研究[J]. 小型微型计算机系统,2000,21(9):940-943.
LU Li-na,CHEN Ya-ping,WEI Heng-yi,et al. Research on the algorithm Apriori mining association rules[J]. Mini-Micro System,2000,21(9):940-943.

[10] 张建明,荣冈. 基于关联规则的故障诊断方法及研究[J]. 化工自动化及仪表,2003,30(5):11-14.
ZHANG Jian-ming,RONG Gang. A method of failure diagnosis based on association rules and study[J]. Control and Instruments in Chemical Industry,2003,30(5):11-14.

[11] 韩建保,罗小江. 基于数据挖掘的坦克传动装置故障征兆识别展望[J]. 车辆与动力技术,2005(4):53-57.
HAN Jian-bao,LUO Xiao-jiang. Perspective of tank transmission failure symptom prediction based on data mining[J]. Vehicle & Power Technology,2005(4):53-57.

[12] 边海燕. 基于数据挖掘技术的动态检测与故障预测分析[D]. 北京:北京化工大学,2004.
BIAN Hai-yan. Dynamic test and fault forecast analysis based on data mining technology[D]. Beijing:Beijing University of Chemical Technology,2004.

[13] 石阳,张红云,马垣. 数据挖掘中关联规则算法及其应用[J]. 鞍山师范学院学报,2002(1):79-81.
SHI Yang,ZHANG Hong-yun,MA Yuan. Association rules algorithms on data mining and its application[J]. Journal of Anshan Teachers College,2002(1):79-81.

[14] 程海明,吴青,赵春华. 油液监测故障诊断关联规则的挖掘研究[J]. 武汉理工大学学报:交通科学与工程版,2004(5):103-105.
CHENG Hai-ming,WU Qing,ZHAO Chun-hua. Associated rules mining in diagnosis based on oil monitoring[J]. Journal of Wuhan University of Technology,2004(5):103-105.

(**责任编辑**: 康鲁豫)

作者简介:

梁志瑞(1959-),男,河北邯郸人,教授,中国电机工程学会城市供电专委会委员,从事电力系统自动化、电气设备参数测量技术及故障分析的教学与研究工作;

陈　鹏(1980-),男,山西临汾人,硕士研究生,研究方向为数据挖掘技术在电厂设备管理系统中的应用(**E-mail**:cp2651@sina.com);

苏海锋(1977-),男,河北栾城人,助教,研究方向为人工智能及其在电力系统中的应用。

Association rule mining in fault monitoring of power plant equipment

LIANG Zhi-rui,CHEN Peng,SU Hai-feng

(North China Electric Power University,Baoding 071003,China)

Abstract: Association rule mining,an important branch of data mining,is to find the inter-dependence among domains,satisfying the given support threshold and confidence threshold,by describing hidden relationships among different data attributes. Parameters at different monitoring points in the power plant are changing with various faults occurred during equipment operation. With the association rule mining algorithm,the relationship between fault phenomenon and fault type during fault occurrence can be set for further fault detection and diagnosis. The concept of the association rule mining and the widely used Apriori algorithm are expatiated. Taking a typical fault of the steam turbine condenser as an example,the feasibility and correctness of the monitoring method are validated by the result analysis.

Key words: association rules; Apriori algorithm; fault monitoring