DOI: 10.19783/j.cnki.pspc.181390

基于改进 Apriori 算法的智能变电站二次设备缺陷关联性分析

陈 勇[1]，李胜男[1]，张 丽[1]，鲁 浩[2]，戴志辉[2]

(1.云南电网有限责任公司电力科学研究院，云南 昆明 650500；2.华北电力大学(保定)，河北 保定 071003)

摘要：智能变电站的出现为大数据的收集、管理提供了技术支持，也为二次设备缺陷数据的关联性分析提供了丰富的数据样本。首先，在此基础上建立了智能变电站二次设备缺陷数据模型。其次，根据智能变电站缺陷数据模型特点对 Apriori 算法进行了改进，降低了算法的时间复杂度和内存占用量。最后，以某市一年的智能变电站二次设备缺陷数据为例，通过改进的 Apriori 算法挖掘缺陷数据各个属性之间的关联性并对关联规则进行了分析。研究表明，该方法能够分析缺陷情况，寻找二次设备薄弱环节，为缺陷巡检方式的制定和检修策略的制定提供支持。与传统 Apriori 算法相比，改进算法的时间复杂度较低。

关键词：Apriori 算法；智能变电站；二次设备；缺陷；关联性分析

Association analysis for defect data of secondary device in smart substations based on improved Apriori algorithm

CHEN Yong[1], LI Shengnan[1], ZHANG Li[1], LU Hao[2], DAI Zhihui[2]

(1. Electric Power Research Institute of Yunnan Power Grid Co., Ltd., Kunming 650500, China;
2. North China Electric Power University (Baoding), Baoding 071003, China)

Abstract: The development of smart substations provide technical support for collection and management of big data, and abundant defect data for related analysis about the secondary device. In view of this, this paper establishes the defect data model of the secondary device in smart substations firstly. Secondly, according to characteristics of the defect data model, the Apriori algorithm is improved to reduce the time complexity and memory occupation. Finally, taking defect data of the secondary device of a city in one year as an example, the association rules of the defect data are acquired and analyzed by the improved Apriori algorithm. The result shows that the proposed method can analyze the device defect, search for the weakness of the secondary device, and provide the support for the formulation of defect inspection modes. Besides, compared with the traditional Apriori algorithm, the improved Apriori algorithm reduces the time complexity.

This work is supported by National Natural Science Foundation of China (No. 51877084), Natural Science Foundation of Hebei Province (No. E2018502063), and Fundamental Research Funds for the Central Universities (No. 2017MS096).

Key words: Apriori algorithm; smart substation; secondary device; defect; association analysis

0 引言

随着经济的发展，通信技术与变电站的联系越来越紧密[1-3]。智能变电站的出现使得二次设备种类多样化、结构复杂化，导致检修工作量的急剧增加和检修人员不足之间的矛盾日益严重[4-6]。IEC61850 主动上送报告服务的应用和光纤通信的发展为变电站缺陷数据的传输提供了很好的基础，为缺陷数据的汇总和综合利用提供了支撑，使缺陷数据关联性分析的实现成为可能[7-9]。另一方面，随着智能变电站各类二次设备的投入运行，相关缺陷数据记录不断积累，将数据挖掘技术引入数据管理系统势在必行。数据挖掘(Data Mining, DM)又称为数据库知识发现(Knowledge Discovery in Databases, KDD)，是一种采用人工智能方式对大量的、随机的、模糊的、不完全的、有噪声的数据进行分析，从而获取有用

基金项目：国家自然科学基金项目资助(51877084)；河北省自然科学基金项目资助(E2018502063)；中央高校基本科研业务费项目资助(2017MS096)

的信息和知识的方法与技术。数据挖掘提取的知识通常可表示为概念、规则、规律、模式等形式，可以被用于信息管理、查询优化、决策支持和过程控制以及数据自身的维护[10-12]。

目前，数据挖掘技术已被广泛应用于解决电力行业中存在的一些问题。例如，文献[13]使用 Apriori 算法挖掘变压器状态参量与状态之间的关联规则以及各状态之间的关联规则，根据关联规则建立基于云－Petri 网的变压器状态分析模型。文献[14]通过数据挖掘技术将变压器状态量分成了单项状态量和综合状态量两类，分别计算单项状态量在故障类型中的常权重系数以及综合状态量的变权重系数，建立了一个较为客观、准确的变压器状态评估体系。

然而，目前关于变电站二次设备缺陷数据的处理还基本停留在对缺陷数据进行简单统计和面向对象的图表呈现阶段，在关联性分析方面的研究并不多见，无法充分挖掘缺陷数据的关联性。文献[15]通过 Apriori 技术对二次设备缺陷数据进行关联性分析，但该方法无法充分挖掘出数据的关联性，只能寻找二次设备薄弱环节和缺陷原因。另外，在建立候选集时没有根据缺陷模型特性对候选集进行筛选，整体算法的时间复杂度和内存占用量都较高。

对于传统 Apriori 算法存在的时间复杂度和内存占用量较高的问题，许多文献也提出了一些改进方案，但都存在或多或少的局限性。文献[16]提出了一种具有动态数据挖掘功能的 Apriori 算法。该算法能在电力系统运维告警系统增添和删除一部分告警记录时，在不用重新扫描数据库的情况下对原关联规则进行修正和补充，但该方法在产生初始关联规则时依旧需要消耗较多的时间。文献[17]提出了基于 Hadoop 算法的数据库分区关联规则算法将数据库进行分区扫描，极大地提高了扫描数据的速度，但是该方法无法挖掘两个分区数据之间的关联性，可能导致部分关联规则的缺失。

为此，本文首先分析智能变电站二次设备缺陷属性，根据属性的特点选取二次设备生产厂家、设备类型、发现方式、缺陷部位和缺陷原因这五个属性构成智能变电站二次设备缺陷数据模型。其次，结合缺陷数据模型对传统 Apriori 算法时间复杂度高的缺点提出两点改进：运用项目识别码减小 Apriori 算法过程中项目候选集数量；在形成频繁项集元素的同时记录包含该频繁项集元素的缺陷记录编号，通过存储的缺陷记录编号计算由该频繁项集连接生成的候选集支持度和置信度。最后，通过算例表明改进 Apriori 算法能根据二次设备部分信息分析二次设备家族性缺陷，帮助检修人员推测缺陷部位、缺陷原因，提供缺陷检修策略，为运维人员确定二次设备薄弱环节和制定巡检方式提供支持。与传统算法相比，改进算法时间复杂度和内存占用量较低。

1 智能变电站二次设备缺陷数据库模型

关联规则(Association Rules)也被称为购物篮分析法(Market Basket analysis)。它是一种能够反映各个事务之间的关联性和相互依存性的规则。关联规则可用来推测一个或几个事务发生之后另外一个事务发生的概率，也可以探索事务之间的潜在联系。

1.1 基本定义

关联规则所涉及的一个事务被称为一个项目(Items)，由不同的项目构成的集合称为项集 I(Itemset)，其元素个数称为项集的长度，长度为 K 的项集称为 K 项集[15]。支持度大于或等于阈值的所有 K 项集的集合就称之为 K 频繁项集，记为 L_K，L_K 中每一个 K 项集称之为 K 频繁项集的一个元素。数据库记为 D，数据库 D 中的第 i 条记录记为 T_i。

1.2 基于关联规则的二次设备缺陷数据库模型

二次设备的运维人员在日常工作中会对所发现和处理的缺陷进行记录和归档，方便日后查看、统计和分析[18]。因此变电站缺陷管理系统中存储着大量的二次设备历史缺陷数据，给变电站二次设备关联性分析提供了数据基础。研究发现目前大部分的缺陷记录都包含多个属性的信息，而这些信息基本可以分为四类，如表 1 所示。

表 1 缺陷属性表

Table 1 Defect attribute table

信息分类	属性
二次设备出厂信息	设备编号、设备类型、专业分类、生产厂家、运行年限、电压等级
所在变电站情况	变电站名称、变电站类别、变电站电压等级
缺陷的基本信息	发现方式、缺陷等级、缺陷部位、缺陷原因、MIS 缺陷情况、缺陷时间、修复时间
人工记录信息	保护异常情况、站内异常情况、处理情况、改进措施

其中，第一类信息记录了二次设备出厂信息。本文提取其中的生产厂家、设备类型两个属性用于后续分析家族性缺陷或某类型设备的薄弱环节。

第二类信息记录了二次设备所在变电站的基本情况。本文不研究缺陷与所处变电站之间的关联性，因此不抽取该属性数据。

第三类信息记录了缺陷的基本情况。发现方式分为监控信号、检修过程、运行巡视、专业巡检四

种。根据发现方式获得的关联规则，检修人员能够针对不同的缺陷制定不同的巡检策略。根据缺陷部位和缺陷原因属性获得的关联规则能在缺陷发生时为检修人员推测缺陷发生部位和原因，甚至为检修人员确定缺陷部位的排查顺序，优先检查关联规则中置信度最高的缺陷部位。而第三类信息中记录的缺陷时间和修复时间偏向于缺陷管理和统计，不抽取该属性数据。因此，本文提取其中的发现方式、缺陷部位、缺陷原因三个属性。

第四类信息运用人类自然语言记录了缺陷发生时保护异常情况、站内异常情况、缺陷处理情况和使用的改进措施。由于目前没有较为成熟的方法能够识别自然语言，抽取有效信息，而利用前三类信息就已经满足二次设备缺陷数据的关联性分析数据要求，因此不抽取该类信息。

据此，本文抽取了生产厂家、设备类型、发现方式、缺陷部位和缺陷原因这五个属性构建数据库，数据库中的部分缺陷记录如表 2 所示。

表 2 部分缺陷记录

Table 2 Some defect records

生产厂家	设备类型	发现方式	缺陷部位	缺陷原因
EXX 公司	F 线路保护	B 监控信号	G 二次回路及辅助设备	H 调试质量不良
EXX 公司	F 母线保护	B 检修过程	G 保护装置本体	H 制造质量不良
EXX 公司	F 故障录波器	B 专业巡视	G 通道传输设备	H 制造质量不良

2 改进的 Apriori 算法

2.1 传统 Apriori 算法

Apriori 算法核心思想是通过 L_{K-1} 的“连接”产生候选集，然后向下封闭检测寻找频繁项集。通过逐层搜索的迭代方法，利用 $K-1$ 项集来搜索 K 项集[14]。“连接”就是当 $K-1$ 频繁项集 L_{K-1} 的两个元素 $L_{K-1,1}=\{I_1,I_2,\cdots,I_{K-2},I_{K-1}\}$ 和 $L_{K-1,2}=\{I_1,I_2,\cdots,I_{K-2},I_K\}(I_{K-1}\neq I_K)$ 只有一项不同时，$L_{K-1,1}$ 和 $L_{K-1,2}$ 能够连接形成 K 候选集的一个元素 $C_{K,1}=\{I_1,I_2,\cdots,I_{K-2},I_{K-1},I_K\}$。

假设频繁项集 I 同时包含项目集 A 和项目集 $B(A\subset I, B\subset I, A\cap B=\Phi)$，定义 $A\Rightarrow B$ 为关联规则 M，则 A 为关联规则 M 的条件，B 为关联规则 M 的结论。关联规则 M 的支持度可写为式(1)。

$$S(A\Rightarrow B)=\mathrm{P}(A\cap B)=\frac{\mathrm{count}(A\cup B)}{\mathrm{count}(D)} \tag{1}$$

式中：count(A)表示在数据库 D 中包含项目集 A 的记录条数；count(B)表示在数据库 D 中包含项目集 B 的记录条数；count($A\cup B$)为在数据库 D 中同时包含 A 和 B 的记录条数；count(D)则表示数据库 D 记录的总数。可见，支持度表示一条记录同时包含 A 和 B 两个项目集的概率。关联规则需满足的支持度的最小阈值称之为最小支持度，记为 $S_{\min}$。

关联规则 M 的置信度 C 可表示为式(2)。

$$C(A\Rightarrow B)=P(A|B)=\frac{\mathrm{count}(A\cup B)}{\mathrm{count}(A)} \tag{2}$$

可见，置信度表示包含 A 项目集的记录包含 B 项目集的概率。关联规则需要满足的置信度的最小阈值称之为最小置信度，记为 $C_{\min}$。

关联规则挖掘是从事物集合中挖掘出满足支持度和置信度最低阈值要求的所有关联规则，这样的关联规则也称强关联规则[19-20]。若两阈值过低，则会包含弱关联性规则；若阈值过高，则会失去部分强关联规则。最小支持度 $S_{\min}$ 是用来衡量关联规则需要满足的最低要求，而最小置信度 $C_{\min}$ 则用来衡量关联规则需要满足的最低可靠性。

Apriori 算法流程如图 1 所示。

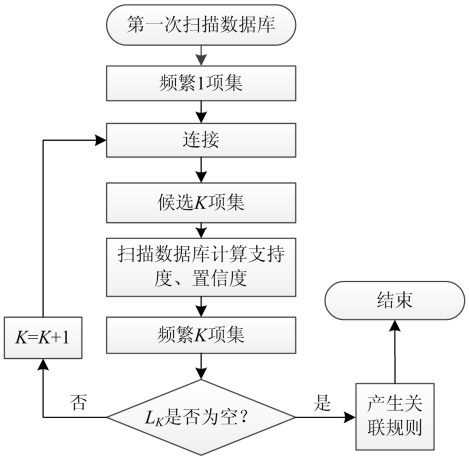

图 1 Apriori 算法流程图

Fig. 1 Apriori algorithm flow diagram

Apriori 算法流程基本步骤如下：

(1) 输入最小支持度 $S_{\min}$ 和最小置信度 $C_{\min}$；

(2) 扫描数据库，从数据库 D 中发现所有的 1 频繁项集，记为 L_1；

(3) 由 L_{K-1}“连接”形成 C_K，遍历计算 C_K 中的每个元素的支持度，删除 C_K 中支持度小于 $S_{\min}$ 的 K 候选集元素，从而获得频繁项集 L_K；

(4) 若 L_K 不为空，则 $K=K+1$，跳转到步骤(3)，L_K 为空时跳转至步骤(5)；

(5) 运用最小置信度 $C_{\min}$ 对频繁项集 L_K 进行筛选，删除 L_K 中置信度小于 $C_{\min}$ 的频繁项集元素之后的集合就是强关联规则。

2.2 改进 Apriori 算法

Apriori 算法具有流程简单易理解、数据要求较

低的优点。然而该算法也有如下缺点：(1) 算法过程中产生大量候选集，算法内存占用量较大；(2) 计算支持度 S 和置信度 C 时需要频繁扫描数据库，算法时间复杂度较高。

针对这些不足，本文对传统 Apriori 算法做了以下两点改进：

(1) 制定项目识别码，删除包含同属性多个项目的候选集元素，减少候选集元素数量。由于传统 Apriori 算法是不分属性地随机将频繁项集“连接”起来产生候选集，经常会出现一个候选集元素中包含同属性的多个项目，例如{厂家 1，监控信号，二次回路及辅助设备}与{厂家 2，监控信号，二次回路及辅助设备}这两个频繁项集“连接”产生候选集{厂家 1，厂家 2，监控信号，二次回路及辅助设备}，这个候选集元素中含有两个厂家项目，而一台设备一般只有一个厂家，因此该候选集是无效的。本文为了识别和删除这类无效候选集，引入了项目属性识别码。

根据 1.2 节智能变电站二次设备缺陷数据模型可知，本文只抽取五个属性的项目，因此只需在生产厂家、设备类型、发现方式、缺陷部位和缺陷原因这五个属性的项目前面分别添加 E、F、B、G、H 这五个识别码，就能识别项目属于那种属性，如表 2 所示。例如，对于属于设备类型的“线路保护”、“母线保护”、“故障录波器”加上识别码 F 变为“F 线路保护”、“F 母线保护”、“F 故障录波器”，就能识别这三种项目均属于设备类型属性。值得注意的是，识别码方法在使用过程中应根据实际缺陷数据模型制定识别码。

根据项目识别码机制，原候选集就会变成{E 厂家 1，E 厂家 2，B 监控信号，G 二次回路及辅助设备}的形式。利用识别码对生成的候选集进行检查，很容易就识别出候选集元素包含了 E 厂家 1、E 厂家 2 这两个同属性项目，从而删除该无效候选集。

该方法能够删除无效候选集，节约内存，减少不必要的频繁项集筛选，降低了算法的时间复杂度和空间复杂度。

(2) 在形成频繁项集元素的同时记录包含该频繁项集元素的缺陷记录编号，通过存储的缺陷记录编号计算由该频繁项集连接生成的候选集的支持度和置信度，减少数据库扫描次数，提高算法效率。

改进 Apriori 算法的支持度可表示为式(3)。

$$S(C_{K,1})=\frac{L_{K-1,1}\text{与}L_{K-1,2}\text{记录编号交集元素个数}}{\text{count}(D)} \tag{3}$$

式中：$C_{K,1}$ 是由 $L_{K-1,1}$ 和 $L_{K-1,2}$ 两个 $K-1$ 频繁项集元素“连接”而成的候选集，规则为 $A\Rightarrow B$，而且 $A\subseteq L_{K-1,1}$，$B\subseteq L_{K-1,2}$。

本文改进 Apriori 算法的置信度可表示为式(4)。

$$C(C_{K,1})=\frac{L_{K-1,1}\text{与}L_{K-1,2}\text{记录编号交集元素个数}}{L_{K-1,1}\text{记录编号元素个数}} \tag{4}$$

下面举例说明该改进算法。表 3 为数据库记录，表 4 和表 5 是用传统 Apriori 算法对表 3 数据计算获得的 1 频繁项集 L_1 和候选集 C_2。

表 3 数据库记录

Table 3 Database project

TID	项目
1	I_1, I_2
2	I_2
3	I_2, I_3
4	I_1, I_2
5	I_1, I_3

表 4 频繁项集 L_1

Table 4 Frequent item set L_1

L_1	支持度	置信度	记录编号
I_1	3/5	100%	1,4,5
I_2	4/5	100%	1,2,3,4
I_3	2/5	100%	3,5

表 5 候选集 C_2

Table 5 Candidate sets C_2

C_2	支持度	置信度	记录编号
I_1, I_2	0.4	66.6%	1,4
I_1, I_3	0.2	33.3%	5
I_2, I_3	0.2	25%	3

用改进的算法计算，步骤如下。在生成频繁项集 L_1 元素时把包含该频繁项集元素的缺陷记录编号存储在“记录编号”属性中，如表 4 所示。此时使用改进算法计算候选集元素{I_1,I_2}的支持度和置信度。由表 4 可见，候选集元素{I_1,I_2}是由 L_1 的元素{I_1}和{I_2}连接而成，查找{I_1}的记录编号为{1,4,5}，{I_2}的记录编号为{1,2,3,4}，这两个记录编号交集为{1,4}，并集为{1,2,3,4,5}。根据式(3)、式(4)计算得{I_1,I_2}的支持度为 S=2/5=0.4，置信度为 C=2/3≈0.666，与表 5 中传统算法获得的结果相等。可见，改进算法可在不扫描原数据库的情况下获得与传统算法相同的支持度与置信度，降低了算法的时间复杂度，提高了算法速率。另外，随着数据库记录的增多，该算法节约时间的效果会随着数据库记录的增多而提升。

3 算例分析

3.1 数据库处理

以某市一年的 264 条变电站二次设备缺陷数据

为例进行关联性分析。首先按照 1.1 节提出的智能站二次设备数据库模型抽取缺陷数据形成二次设备缺陷数据库。

由于智能变电站二次设备具备高可靠性，二次设备缺陷类型具备分散性。在此情况下，盲目删除部分属性缺失的记录会缩小样本，易丢失一些小样本包含的关联规则，因此本文不删除部分属性缺失的记录，而是通过查询检修报告、能量管理系统等方法补全缺失数据。

经统计，在该缺陷数据库中，共有生产厂家 24 个，设备类型有“线路保护”、“主变保护”、“合并单元”、“故障录波器”等 13 种类型；发现方式有“监控信号”、“专业巡视”、“运行巡视”、“检修过程”4 种类型；缺陷部位包括“二次回路及辅助设备”、“保护装置本体”、“通道传输设备”、“保护通道及接口设备”等 7 种类型；缺陷原因包括“调试质量不良”、“制造质量不良”、“设备老化”、“运行维护不良”、“其他缺陷原因分类”这 5 种类型。

3.2 用改进算法产生强关联规则

本文期望挖掘出生产厂家、设备类型、发现方式、缺陷部位和缺陷原因这五个属性之间的关联性，因此抽取这五个属性的数据构成缺陷数据库，并在过程中加入项目识别码表明项目所属的属性。每一个频繁项集元素中不能包含同属性的多个项目，因此最大的频繁项集就是 5 频繁项集，Apriori 算法最多获得 5 频繁项集。

在算例的数据库中，各类缺陷所占比例较小，在 264 条记录中某一种缺陷可能只有三到四条记录，若设置过高的最小支持度 S_{min} 和最小置信度 C_{min} 容易失去一部分的强关联规则。因此，设置最小支持度值 S_{min}=1.1%，最小置信度 C_{min}=40%。首先扫描数据库获得 1 频繁项集，并记录下包含各个频繁项集的记录编号。接着运用改进的 Apriori 算法以及式(3)和式(4)计算各个频繁项集元素的支持度和置信度，筛选出支持度大于 S_{min} 的所有频繁项集。最后，筛选出 2 频繁项集、3 频繁项集、4 频繁项集和 5 频繁项集中置信度大于 C_{min} 的频繁项集。

经对该市一年的变电站二次设备缺陷数据关联性分析，一共获得 2 个项目的强关联规则 92 个，3 个项目的强关联规则 87 个，4 个项目的强关联规则 31 个，5 个项目的强关联规则 1 个。由于本文样本较少，一种缺陷情况可能出现多次，或没有类似的缺陷情况，因此置信度偏高。

部分关联规则结果如表 6 所示。从表 6 所展示的部分强关联规则中可以得出如下结论：

(1) 关联规则能够根据已知条件推测缺陷所在部位。在关联规则 1 中，当条件为生产厂家 1、监控信号时，结论为缺陷部位在保护装置本体关联规则置信度为 75%。在关联规则 4 中，当条件为生产厂家 3、故障录波器、监控信号时，结论为缺陷部位在通道传输设备的关联规则置信度为100%。检修人员在收到缺陷报警信号之后，将如关联规则 1 和 4 这种类型的关联规则与报警设备的厂家、设备类型、发现方式等现场信息相结合，推测该设备缺陷发生的部位。置信度越高的规则推测的缺陷发生部位越准确可靠。检修人员可以优先检查由置信度高的关联规则推测的缺陷部位。

表 6 部分强关联规则

Table 6 Strong association rules

编号	关联规则	支持度	置信度
1	E 厂家 1，B 监控信号⇒G 保护装置本体	5.11%	75%
2	E 生产厂家 2，B 监控信号，G 保护装置本体⇒H 制造质量不良	2.84%	100%
3	E 生产厂家 1，F 线路保护，G 保护装置本体⇒B 监控信号	2.27%	100%
4	E 生产厂家 3，F 故障录波器，B 监控信号⇒G 通道传输设备	1.70%	100%
5	E 生产厂家 4，F 线路保护，B 监控信号，G 保护装置本体⇒H 制造质量不良	3.41%	46.15%
6	E 生产厂家 5⇒H 制造不良	14.2%	55.56%
7	E 生产厂家 6⇒G 通道传输设备	17%	100%

(2) 关联规则能够根据已知条件推测设备发生缺陷的原因。在关联规则 2 中，当条件为生产厂家 2、监控信号、保护装置本体时，结论为制造不良的置信度为 100%。在关联规则 5 中，当条件为生产厂家 4、线路保护、监控信号、保护装置本体时，结论为制造不良的缺陷原因的置信度为 46.15%。检修人员在收到缺陷报警信号之后，将如关联规则 2 和 5 这种类型的关联规则与报警设备的生产厂家、设备类型、发现方式和缺陷部位等现场信息相结合，推测缺陷原因。置信度越高的规则推测的缺陷原因也就越准确可靠。由此可见，关联规则可以推测缺陷原因，提高检修人员排查缺陷的效率，缩短修理缺陷的时间。

(3) 关联规则能够根据已知条件为巡检方式的

制定提供支持。在关联规则 3 中，当条件为生产厂家 1、线路保护、保护装置本体时，结论为监控信号的置信度为 100%。也就是说，当厂家 1 生产的线路保护设备在保护装置本体处存在缺陷时，该缺陷基本都是通过监控信号发现的。对于该缺陷，检修人员可以有针对性地增强监控能力，消除这一类缺陷。这一类的关联规则能够呈现条件与发现方式之间的强关联性，作为运维人员安排何种运维排查缺陷方式的依据。如果某个缺陷发生较频繁，运维人员可根据该缺陷的关联规则有针对性地加强某一种运维排查缺陷方式来及时发现这一类缺陷。另外，当某一关联规则的结论是“检修过程”这一发现方式，并且该设备缺陷发生较频繁时，检修人员可以考虑缩短该设备的检修周期，提高检修频率来及时发现该设备缺陷。

(4) 关联规则能够发现二次设备的家族性缺陷。在关联规则 6 中，生产厂家 5 与制造不良具有强关联性，置信度为 55.56%。可见生产厂家 5 生产的设备极有可能存在制造不良的问题，应建议该厂家提升制造工艺，提高设备出厂验收要求。

(5) 关联规则能够发现二次设备薄弱环节。在规则 4 中，当条件为生产厂家 3、故障录波器、监控信号时，结论为缺陷部位在通道传输装置，置信度为 100%。在规则 7 中，当条件为生产厂家 6 时，结论为缺陷部位在通道传输装置，置信度也是 100%。可见生产厂家 3 的故障录波器设备薄弱环节在通道传输装置，生产厂家 6 的设备经常会发生通道传输装置缺陷告警。这一类的关联规则能够给检修人员提供二次设备薄弱环节的信息，帮助检修人员采取措施来加强二次设备薄弱环节的检测或者消除某一类设备的薄弱环节。

3.3 改进算法的时间复杂度

本文将 500 条、1 000 条和 5 000 条记录的 5 列随机生成数组作为算法时间复杂度的研究样本。运用传统 Apriori 算法和改进 Apriori 算法分别对不同记录总数的样本进行关联性分析，最小支持度取 $S_{\min}=4/\mathrm{count}(D)$，统计不同算法所需时间，结果如表 7 所示。

表 7 时间复杂度比较表

Table 7 Time complexity comparison table

记录条数	传统 Apriori 算法	改进 Apriori 算法	时间差
500	48.206 4 s	18.590 2 s	29.616 2 s
1 000	146.168 2 s	28.785 5 s	117.382 7 s
5 000	3 330.597 5 s	377.903 3 s	2 952.694 2 s

从表 7 可见，改进 Apriori 算法时间复杂度较低、算法效率较高：

(1) 在相同样本情况下，改进 Apriori 算法所需时间小于传统 Apriori 算法，即改进 Apriori 算法的时间复杂度较传统 Apriori 算法低。

(2) 随着记录条数的增加，传统 Apriori 算法和改进 Apriori 算法所需时间差将增大。即样本越大，改进 Apriori 算法节约时间越多。

4 结论

本文提出了智能变电站二次设备缺陷数据库模型，通过改进的 Apriori 算法对智能变电站二次设备缺陷数据进行数据挖掘获得关联规则，能够有效提高算法速度、降低内存要求，获得更加有效的关联规则。缺陷数据的关联规则能够用于分析家族性缺陷，寻找二次设备的缺陷薄弱环节，在排查缺陷部位、缺陷原因时为检修人员提供建议。

但是，本文也没有提出能够识别和合并相似关联规则的方法，无法智能选择最合适的支持度和置信度的阈值，这些方面有待进一步研究。

参考文献

[1] 高翔, 张沛超, 章坚民. 电网故障信息系统应用技术[M]. 北京: 中国电力出版社, 2007: 70-185.

[2] 戴志辉, 张天宇, 刘譞, 等. 面向状态检修的智能变电站保护系统可靠性分析[J]. 电力系统保护与控制, 2016, 44(16): 14-21.
DAI Zhihui, ZHANG Tianyu, LIU Xuan, et al. Research on smart substation protection system reliability for condition-based maintenance[J]. Power System Protection and Control, 2016, 44(16): 14-21.

[3] APOSTOLOV A. Efficient maintenance testing in digital substations based on IEC 61850 edition 2[J]. Protection and Control of Modern Power Systems, 2017, 2(2): 407-420. DOI: 10.1186/s41601-017-0054-0.

[4] 张友强, 王洪彬, 刁兴华, 等. 计及保护失效的智能变电站二次系统综合风险评估研究[J]. 电力系统保护与控制, 2018, 46(22): 155-163.
ZHANG Youqiang, WANG Hongbin, DIAO Xinghua, et al. Integrated risk assessment of intelligent substation secondary system considering the protection failure[J]. Power System Protection and Control, 2018, 46(22): 155-163.

[5] 郭采珊, 蔡泽祥, 潘天亮, 等. 基于信息可达性的智能变电站继电保护系统风险评估方法[J]. 电网技术, 2018, 42(9): 3041-3048.
GUO Caishan, CAI Zexiang, PAN Tianliang, et al. Risk assessment for protection system in smart substation considering information reachability[J]. Power System Technology, 2018, 42(9): 3041-3048.

[6] 孙金莉, 李煜磊, 冯凝, 等. 智能变电站二次设备缺陷分析专家系统的研究与应用[J]. 电网与清洁能源, 2016, 32(10): 95-98.

SUN Jinli, LI Yulei, FENG Ning, et al. Research and application of the expert system for defect analysis of the secondary equipment in smart substations[J]. Power System and Clean Energy, 2016, 32(10): 95-98.

[7] 戴志辉, 谢军, 陈曦, 等. 基于动态贝叶斯网络的智能变电站监控系统可靠性分析[J]. 电力系统保护与控制, 2018, 46(23): 68-76.
DAI Zhihui, XIE Jun, CHEN Xi, et al. Dynamic Bayesian network based reliability evaluation of supervision and control system in smart substations[J]. Power System Protection and Control, 2018, 46(23): 68-76.

[8] 刘洋, 马进, 张籍, 等. 考虑继电保护系统的新一代智能变电站可靠性评估[J]. 电力系统保护与控制, 2017, 45(8): 147-154.
LIU Yang, MA Jin, ZHANG Ji, et al. Reliability evaluation of a new generation smart substation considering relay protection system[J]. Power System Protection and Control, 2017, 45(8): 147-154.

[9] 王同文, 谢民, 孙月琴, 等. 智能变电站继电保护系统可靠性分析[J]. 电力系统保护与控制, 2015, 43(6): 58-66.
WANG Tongwen, XIE Min, SUN Yueqin, et al. Analysis of reliability for relay protection systems in smart substation[J]. Power System Protection and Control, 2015, 43(6): 58-66.

[10] 李勋, 龚庆武, 杨群瑛, 等. 基于数据挖掘技术的保护设备故障信息管理与分析系统[J]. 电力自动化设备, 2011, 31(9): 88-91.
LI Xun, GONG Qingwu, YANG Qunying, et al. Fault information management and analysis system based on data mining technology for relay protection[J]. Electric Power Automation Equipment, 2011, 31(9): 88-91.

[11] 俞京锋, 阮远峰, 于乐, 等. 基于 Hadoop 分布式并行计算在继电保护整定计算中的应用[J]. 电网与清洁能源, 2014, 30(12): 96-108.
YU Jingfeng, RUAN Yuanfeng, YU Le, et al. Application of relay protection setting calculation based on Hadoop distributed parallel technology[J]. Power System and Clean Energy, 2014, 30(12): 96-108.

[12] 周云祥. 基于数据挖掘的变电站监控后台告警信号自动分析[D]. 北京: 华北电力大学, 2016.
ZHOU Yunxiang. Automatic analysis substation monitoring background alarm signals based on data mining[D]. Beijing: North China Electric Power University, 2016.

[13] 王有元, 周立玮, 梁玄鸿, 等. 基于关联规则分析的电力变压器故障马尔科夫预测模型[J]. 高电压技术, 2018, 44(4): 1051-1058.
WANG Youyuan, ZHOU Liwei, LIANG Xuanhong, et al. Markov forecasting model of power transformer fault based on association rules analysis[J]. High Voltage Engineering, 2018, 44(4): 1051-1058.

[14] 李黎, 张登, 谢龙君, 等. 采用关联规则综合分析和变权重系数的电力变压器状态评估方法[J]. 中国电机工程学报, 2013, 33(24): 152-159.
LI Li, ZHANG Deng, XIE Longjun, et al. A condition assessment method of power transformers based on association rules and variable weight coefficients[J]. Proceedings of the CSEE, 2013, 33(24): 152-159.

[15] 张延旭, 胡春潮, 黄曙, 等. 基于 Apriori 算法的二次设备缺陷数据挖掘与分析方法[J]. 电力系统自动化, 2017, 41(19): 147-151.
ZHANG Yanxu, HU Chunchao, HUANG Shu, et al. Apriori algorithm based data mining and analysis method for secondary device defect[J]. Automation of Electric Power Systems, 2017, 41(19): 147-151.

[16] 刘雪芬. 基于电力运维告警数据的诊断系统的设计与实现[D]. 北京: 华北电力大学, 2016.
LIU Xuefen. Design and implementation of the diagnosis system based on the alarm data of electric power operation and maintenance[D]. Beijing: North China Electric Power University, 2016.

[17] 李若晨. 基于并行的 Apriori 数据挖掘算法的研究[D]. 吉林: 吉林大学, 2017.
LI Ruochen. Research on a parallel data mining algorithm Apriori[D]. Jilin: Jilin University, 2017.

[18] 蓝鹏昊. 基于变电站运行信息的智能数据挖掘[D]. 广州: 华南理工大学, 2013.
LAN Penghao. Intelligent data mining based on running information of substations[D]. Guangzhou: South China University of Technology, 2013.

[19] 王磊, 陈青, 高洪雨, 等. 基于大数据挖掘技术的智能变电站故障追踪构架[J]. 电力系统自动化, 2018, 42(3): 84-91.
WANG Lei, CHEN Qing, GAO Hongyu, et al. Framework of fault trace for smart substation based on big data mining technology[J]. Automation of Electric Power Systems, 2018, 42(3): 84-91.

[20] 陈碧云, 丁晋, 陈绍南, 等. 基于关联规则挖掘的电力生产安全事故事件关键诱因筛选[J]. 电力自动化设备, 2018, 38(4): 68-74.
CHEN Biyun, DING Jin, CHEN Shaonan, et al. Selection of key incentives for power production safety accidents based on association rule mining[J]. Electric Power Automation Equipment, 2018, 38(4): 68-74.

收稿日期：2018-11-06； 修回日期：2019-01-03

作者简介：

陈 勇(1977—)，男，硕士，高级工程师，研究方向为继电保护；

李胜男(1974—)，女，硕士，高级工程师，研究方向为继电保护；

张 丽(1988—)，女，学士，高级工程师，研究方向为继电保护。

(编辑 魏小丽)