关联规则在输电网安全性评价中的应用

程　政[1],雷　霞[2],柏小丽[1],徐博海[3]
(1. 泸州电业局,四川 泸州　646000; 2. 西华大学电气信息学院,四川 成都　610039;
3. 国电大渡河瀑布沟水利发电总厂,四川 雅安　625304)

摘　要: 一种新的查找方法:数据挖掘方法,用来查找输电网存在的危险点。建立了一个智能化的数据挖掘工具,从输电网安全性评价数据库的大量历史数据中获取输电网中存在的危险点从而指导管理者进行决策。数据挖掘工具的核心是:采用 Apriori 算法找关联规则从而判断影响输电网安全性的危险点。通过危险点查找,确认发生事故的可能性及其严重程度,提出相应的整改和控制措施,达到预防、控制隐患和事故。

关键词: 数据挖掘; 关联规则; 危险点; Apriori 算法

Abstract: An entirely new searching method, data mining, has been put forward to figure out the dangerous points that might exist in transmission network. An intelligent data mining tool is established, and on the basis of numerous historical data in safety evaluation database of transmission network, the data mining method can obtain the dangerous points in transmission network to guide the administrators to make correct decisions. The core of data mining tool is to find association rules through Apriori algorithm to judge the dangerous points that may affect the security of transmission network. By searching for the dangerous points, the possibility and severity of the accident can be confirmed, and then the relevant amendments and control measures are proposed to prevent and control the safety hazards and accidents.

Key words: data mining; association rules; dangerous point; Apriori algorithm
中图分类号: TM762　**文献标志码:** A　**文章编号:** 1003 - 6954(2011)04 - 0052 - 03

0　引　言

安全性评价属于风险管理范畴,是预防和控制企业事故行之有效的方法[1]。安全性评价是度量、预测系统安全基础、控制事故的重要措施。中国已用法律形式将“安全第一,预防为主”确定为劳动保护方针,也是电力安全生产和建设管理的基本方针。编制安全性评价标准就是要切实贯彻“安全第一,预防为主”的方针。针对电网运行、设备工况、生产环境、作业过程等进行安全性评价实现对事故的超前预测和控制,达到消灭和减少事故的目的。1990 年,华北电网公司借鉴国外风险评估等现代安全管理办法,开始了发电机组并网安全性评价,目前已在全国全面推开。随着安全性评价在电力系统中的应用,一些安全性评价管理系统已投入实际应用,但仅仅完成了安评数据的统计以及个别数据的追踪功能,对于查评中隐藏的信息没有进行有效的分析和挖掘。如何处理这些安评数据成为研究的首要问题,数据挖掘技术应运而生。数据挖掘也称数据库知识发现,它从大量的、不完全的、有噪声的、随机的实际应用数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[2]。下面建立了数据挖掘模型,挖掘影响输电网安全的危险点,从而指导管理者开展输电网安全性评价工作。

1　基本原理

1.1　关联规则基本概念

关联规则反映一个事物与其他事物之间的相互依存性和关联性。如果两个事物或者多个事物之间存在一定的关联关系,那么,其中一个事物就能通过其他事物预测到。一般来说,关联规则就是描述数据库中数据项(属性、变量)之间所存在的潜在关系的规则。设 $I=\{i1,i2,\cdots,im\}$ 是 m 个不同项目的集合,D 是针对 I 事物的集合,每一笔事物包含若干项目 $i1,i2,\cdots,ik\in I$。关联规则形如 $X\Rightarrow Y$ 的蕴含式,其中 $X\subset I, Y\subset I, X\cap Y=\Phi$。关联规则 $X\Rightarrow Y$ 在事务集 D

中成立，具有两个规则兴趣度度量 - 支持度和置信度，它们分别反映发现规则的有用性和确定性。

定义 1　支持度　X 的支持度是事物集中 A 出现的事物数与总的事物数之比，即

$$\text{support}(X) = \frac{\text{support\_count}(X)}{\text{support\_count}(D)} \quad (1)$$

定义 2　置信度　规则 $X \Rightarrow Y$ 的置信度是事物集中 X、Y 同时出现的事物数与 X 出现的事物数之比，即

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{support\_count}(X \cup Y)}{\text{support\_count}(X)} \quad (2)$$

如果规则的支持度大于最小支持度则认为此规则是频繁项集，否则为非频繁项集。同时满足最小支持度与最小可信度两属性的规则称为强关联规则。关联规则挖掘的目的就是从数据库中挖掘出满足用户要求的最小支持度与最小可信度的强关联规则。

1.2　关联规则挖掘一般步骤

挖掘关联规则问题一般可以分解为以下两个子问题[9]。

(1) 找出存在于事物数据库中的所有频繁项集，即找出所有支持度满足用户所规定的最小支持度阈值的项集。

(2) 用频繁项集生成候选关联规则，然后验证候选关联规则是否满足用户所规定的最小可信度阈值。若满足，该候选关联规则为要找的关联规则。

2　关联规则实现过程

2.1　关联规则模型的建立

要处理的问题是如何从数据源中挖掘到想要的危险点。那么建立了图 1 给出了关联规则模型。

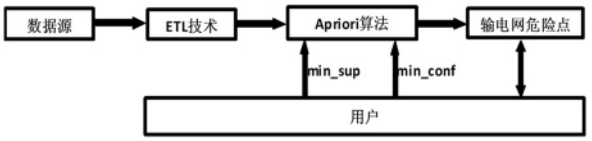

图 1　关联规则模型

模型解释：数据源选择是输电网安全性评价管理系统中的数据；ETL 技术指的是对数据源中数据的清理、转换等；应用 Apriori 算法时用户需对算法中的参数进行设置，其中 min_sup 代表最小支持度，min_conf 代表最小置信度。通过数据挖掘技术挖掘出的危险点，可以直观地展现给决策者。

2.2　模型的求解

2.2.1　数据预处理

由于自查评表中数据比较完整、数据易处理的。下面对数据进行如下处理。

(1) 问题严重程度缺失的处理

在自查评过程中，问题严重程度的缺失是由于用户在填写自查评结果时漏填，因此为了不遗漏任何危险的因素，默认为问题的严重程度为严重。

(2) 数据错误的处理

数据错误产生的原因主要有两个：设计数据时没有进行约束；数据的人为输入错误。前者是在系统设计时没有对用户的输入进行约束，使得用户可以输入不满足要求的数据。后者是由于很多数据都是以字符串的形式来存储的，无法使用约束来保证数据的正确性，而且由于用户知识水平和文化背景的差异，输入时往往会拼写出错或者录入完全错误的数据。

在自查过程中，由于参评人员对标准的理解有差异，打出的分数不满足系统要求，甚至可能打出错误的分数，因此得分率就有可能大于 1 或者小于 0，必须进行清理。首先通过 SQL 语句找到所有错误的得分率，将这些得分率都默认为 0。

(3) 数据转换

由于自查评表中的得分率在 [0 - 1] 区间，问题的严重程度分为一般和严重两种。所用的 Apriori 算法是基于布尔型关联规则的挖掘，那么现在将数据离散化处理。将得分率记为 selfrate，问题的严重程度记为 plevel。得分率在区间 [0 - 0.5] 之间记为 A1；在 [0.5 - 1] 之间记为 A2。问题一般记为 B1；问题严重记为 B2。那么自查评表就转化为最终的数据表，如表 1 所示。

表 1　关联规则模型最终事务表

tid	selfrate	plevel
1	A1	B1
2	A2	B2
3	A2	B1
…	…	…

2.2.2　数据挖掘的实现

首先采用 Apriori 算法生成频繁项集，然后由频繁项集根据最小支持度和最小置信度产生强规则。基于 Apriori 算法的数据挖掘流程如图 2 所示。

2.2.3　关联规则在输电网安全性评价中的应用

输电网自查评表保存着输电网安全性评价中用户自查评时产生的数据，其中包括查评项目、查评得

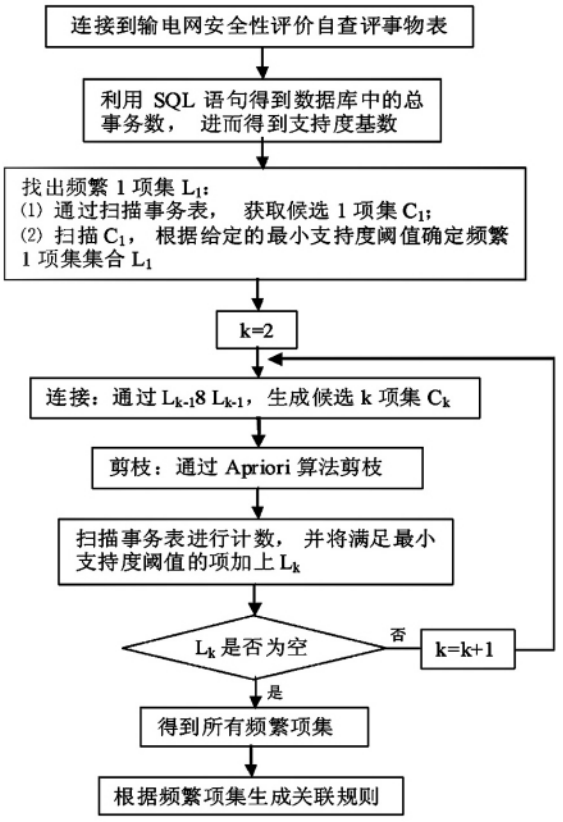

图 2　数据挖掘流程

分、得分率等等。以自查评的项目为挖掘对象，以自查评表中的得分率(selfrate)和问题严重程度(plevel)为研究对象。通过前 200 次查评的历史数据进行挖掘，对得分率和问题严重程度之间的关联规则来判断输电网安全性评价指标是否存在危险点。

对于《输电网安全性评价标准》中查评项目 2.2.1[5]，国家标准是这样描述的：主力大容量电厂是否介入本网最高一级电压电网。评分标准及方法：符合要求的满分，基本符合要求得 60% 标准分，不符合要求不得分。对于次项目，取之前 200 次的查评数据进行分析，通过数据挖掘找到得分率和问题严重程度之间的关联规则，对于得分率低且问题严重的关联规则，可以得出项目 2.2.1 是存在危险点的。

2.2.4　挖掘结果及解释

本模型对自查评表中查评项目 2.2.1 项前 200 次的查评历史数据进行分析，设定最小支持度为 0.5，最小置信度为 0.6。最后得到 3 条强关联规则。例如规则：A1⇒B2，其中支持度为 50.2%，置信度为 67.5%。意味着当“得分率”在 [0 – 0.5] 之间时，“问题严重程度”的概率为 67.5%。那么针对这种得分率低而且问题严重的关联规则，管理者可以对 2.2.1 项提前提出控制措施，达到预防事故的目的。

3　结　语

介绍了关联规则在输电网安全性评价中的应用，并建立了关联规则模型，挖掘输电网中存在的危险点。针对危险点，对输电网提出相应的整改措施，对输电网安全性评价有一定的指导作用。

参考文献

[1] JiaweiHan, MichelineKambe 著，范明，孟小峰译. 数据挖掘概念与技术 [M]. 北京：机械工业出版社，2007.
[2] 王金萍，樊凤林，刘发旺，等. 安全性评价在电力企业中的应用研究 [J]. 华北电力技术，2005(5)：23 – 26.
[3] 国家电网公司. 供电企业安全性评价标准 [S]. 北京：中国电力出版社，2002.
[4] 骆嘉伟，彭蔓蔓，陈景燕，等. 基于消费行为的 Apriori 算法的研究 [J]. 计算机工程，2003，29(5)：72 – 74.
[5] 杨辅祥，刘云超，段智华. 数据清理综述 [J]. 计算机应用研究，2002，19(3)：3 – 5.
[6] 高艳霞. Apriori 算法在学生成绩管理中的应用 [J]. 计算机时代，2009(8)：30 – 31.
[7] 陶建江，张文献. 关联规则挖掘的基本算法 [J]. 计算机工程，2004，15(30)：34 – 35.
[8] 高杰，理绍军，钱锋，数据挖掘中关联规则算法的研究及应用 [J]. 2006(36)：128 – 131.
[9] 程政，雷霞，廖翔，等. 数据挖掘在电网安全性评价中的应用 [J]. 2010(8)：97 – 99.

作者简介：

程　政(1986)，男，硕士研究生，研究方向为计算机技术在电力系统中的应用；

雷　霞(1973)，女，四川南充人，博士，硕士生导师，研究方向为配电自动化和电力市场；

柏小丽(1985)，女，四川达州人，硕士，研究方向为电力市场及配电自动化；

徐博海(1984)，男，河北唐山人，本科，研究方向为电厂运行维护。

(收稿日期：2011 – 05 – 30)