

面向智能电网大数据关联规则挖掘的频繁模式网络模型

孙丰杰,王承民,谢 宁

(上海交通大学 电子信息与电气工程学院,上海 200240)

摘要:针对目前关联规则挖掘频繁树(FP-Tree)算法实现较困难以及难以处理数据库更新的缺点,提出了频繁模式网络(FP-network)模型,将关联规则挖掘所需要的信息压缩到一个无向网络图上,并建立事务-项目关联矩阵,从而进行数据存储和数据挖掘。FP-network模型适用于智能电网大数据的关联规则挖掘。以关联规则挖掘在输电线路故障分析领域的应用为例进行算例分析,结果表明所提出的FP-network关联规则挖掘算法不仅继承了FP-Tree算法的优点,而且只需扫描一次数据库,也便于数据库的维护和更新,从而提高了智能电网大数据关联规则挖掘的效率。

关键词:智能电网;大数据;关联规则;FP-tree算法;FP-network算法

中图分类号:TM 761

文献标识码:A

DOI:10.16081/j.issn.1006-6047.2018.05.016

0 引言

智能电网是利用现代网络信息技术等实现电网设备间的数据信息交换,从而实现电网实时自动化控制、智能调节、在线决策分析等功能新型电网^[1]。

智能电网的建设积累了海量数据资源,目前电力公司“用数据管理企业,用信息驱动业务”的需求日益迫切。而智能电网大数据具有4V特征,即数量大(Volume)、种类多(Variety)、速度快(Velocity)、价值密度低(Value)。传统的数据处理方法已经难以满足需求,因此学者们提出了一系列大数据挖掘算法。

关联规则挖掘算法由Agrawal等在文献[2]中首先提出,该方法从大量历史数据中寻找频繁项或属性之间的关联性。现有的关联规则挖掘方法主要是Apriori算法及频繁模式树FP-Tree(Frequent Pattern-Tree)算法^[3-4]。Apriori算法的主要缺点是需要寻找大量的候选项目集,当数据库较大时,存在组合爆炸问题,同时,Apriori算法需要多次扫描数据库,增加了计算的负担。

针对Apriori算法的缺点,J. Han提出了利用FP-Tree产生频繁项集的方法^[5-6]。FP-Tree算法将提供频繁项集的数据库压缩到FP-Tree上,然后从初始后缀模式开始,构造条件模式基,再形成条件FP-Tree,并递归地在该树上进行挖掘,其主要优点体现在:不需要产生候选项,仅需要构造FP-Tree和条件FP-Tree,通过递归地访问FP-Tree产生频繁模式;对事务数据库仅需2次遍历,第1次遍历产生频繁1-项集,第2次遍历用于创建FP-Tree,从而极大

地降低了访问数据库的次数。FP-Tree算法存在的主要问题为:挖掘过程中需要不断递归地生成“树”,增加了时空复杂度;FP-Tree和条件FP-Tree每次都需要双向遍历数据库,因此难以处理数据库更新、维护问题。

鉴于目前FP-Tree关联规则挖掘算法存在的问题,本文将提供频繁项集的数据压缩到FP-network上,通过形成关联矩阵,进行计算机存储和挖掘。本文方法继承了FP-Tree模型无需产生候选项以及不重复扫描数据库的优点,同时克服了FP-Tree模型生成复杂及更新、维护困难的缺点,特别适用于智能电网大型数据库,挖掘智能电网中的复杂规律。

1 FP-network 模型

电力系统的数据库通常是事务和项目之间的关联,如表1所示的事务数据库,其中第1列为事务编号,第2列为项目集合,即事务包含哪些项目,项目集合为 $\{I_1, I_2, I_3, I_4, I_5\}$ 。

表1 事务数据库列表

Table 1 Transaction database table

事务	项目集
T_{001}	$\{I_1, I_2, I_5\}$
T_{002}	$\{I_2, I_4\}$
T_{003}	$\{I_2, I_3\}$
T_{004}	$\{I_1, I_2, I_4\}$
T_{005}	$\{I_1, I_2, I_3\}$
T_{006}	$\{I_2, I_3\}$
T_{007}	$\{I_1, I_3\}$
T_{008}	$\{I_1, I_2, I_3, I_5\}$
T_{009}	$\{I_1, I_2, I_3\}$

1.1 FP-network 模型的网络图形式

根据表1,建立传统的FP-Tree模型如图1所示。由图1可见,树的生成过程十分复杂。此外,图1所示的FP-Tree模型将提供频繁项的数据库压缩到一个有向树状图上,所以存在维护、更新困难的缺

收稿日期:2017-11-08;修回日期:2018-03-28

基金项目:国家自然科学基金资助项目(51777121)

Project supported by the National Natural Science Foundation of China(51777121)

点。为了避免这些缺点,本文提出了无向的 FP-network 模型。

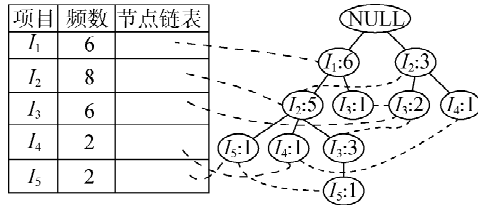


图1 FP-Tree 示意图

Fig.1 Schematic diagram of FP-Tree

首先定义以下概念。

a. 弧容量:扫描事务数据库,第 i 条弧出现的次数,记作 a_i 。

b. 节点频数:扫描事务数据库,节点 j 出现的次数,记为 f_j 。

c. 节点负容量:扫描每条事务,节点 j 最后被扫描的次数,记为 n_j 。

以表 1 为例,建立 FP-network 模型的过程为如下。

a. 将各个项目作为网络中的节点,将每个事物作为网络中的路径。

b. 扫描数据库,事务 T_{001} 包含 3 个项目,因此对应节点频数 f_1 、 f_2 和 f_3 加 1; T_{001} 可以认为是由 2 条弧组成的,即 $\widehat{I_1 I_2}$ 和 $\widehat{I_2 I_5}$,因此 a_1 和 a_2 分别加 1;且 T_{001} 最后扫描 I_5 节点,所以 n_5 加-1。

c. 按照上述原则依次扫描其他事务,所建立的 FP-network 如图 2 所示。

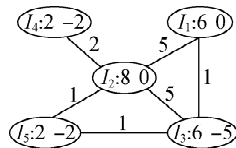


图2 FP-network 示意图

Fig.2 Schematic diagram of FP-network

图 2 所示的 FP-network 有以下特点。

a. 不同于 FP-Tree, FP-network 是无向图,且同一项目对应图中唯一一节点(FP-Tree 中同一项目可能对应多个节点)。

b. 弧容量之和等于所有节点频数与节点负容量的和,即:

$$\sum_{i=1}^n a_i = \sum_{j=1}^m f_j + \sum_{j=1}^m n_j \quad (1)$$

其中, n 为弧的数目; m 为节点数目。

c. FP-network 可能会将某些事务数量扩大。如对于项目 I_5 而言,有 2 个事务与其相关联,分别为 $\{I_1, I_2, I_3, I_5\}$ 和 $\{I_1, I_2, I_5\}$ 。但是图 2 中,节点 I_5 可以找到 4 条路径,分别为 $\{I_1, I_2, I_3, I_5\}$ 、 $\{I_1, I_2, I_5\}$ 、

$\{I_1, I_3, I_5\}$ 、 $\{I_2, I_5\}$,而后 2 条路径实际上并不存在。

1.2 FP-network 模型的矩阵形式

为了避免上述网络图形式的缺点,FP-network 的计算机存储采取事务(路径)-项目(节点)关联矩阵表示方式,即 $T=f(B, I)$ 。仍以表 1 为例,有:

$$B = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}, I = \begin{bmatrix} I_1:6 & 0 \\ I_2:8 & 0 \\ I_3:6 & -5 \\ I_4:2 & -2 \\ I_5:2 & -2 \end{bmatrix}$$

其中, T 为事务集合; I 为项目集合;矩阵 B 为事务-项目关联矩阵,其元素 b_{ij} ($i=1, 2, \dots, 9; j=1, 2, \dots, 5$) 可以定义为:对于事务 i ,如果与项目 j 相关联,则 $b_{ij}=1$,否则 $b_{ij}=0$ 。对于大数据而言,通常事务数目远大于项目数目,因此生成关联矩阵的时间复杂度近似为 $O(\text{事务数目})$,并且对数据库的存储可以转换为存储矩阵布尔矩阵 B 和 I ,极大地节省了内存。

1.3 FP-network 算法步骤

利用 FP-network 算法可以方便地实现关联规则的挖掘,具体步骤如下。

a. 给定最小支持度阈值 $S_{\min}$ 。

b. 扫描数据库,如果 $f_j < S_{\min}$,则删去第 j 个节点信息,构建矩阵 B, I 。

c. 从 n_k 为非零的第 k 个节点开始挖掘。

d. 只保留矩阵 B 中 $b_{ik}=1$ 的节点 k 所有路径集合 $I(I=\{i | b_{ik}=1\})$,仅保留节点 k 之前的节点信息,形成新的矩阵 B, I 。

e. 如果 $|n_k| \geq S_{\min}$,则节点 k 作为一个频繁项集的元素之一,否则不进行任何处理。删去此节点信息,形成新的矩阵 B, I 。转到步骤 b,此过程持续至所有 n_k 为非零的节点挖掘完为止。

如表 1 示例,取 $S_{\min}$ 为 2,因 I_1-I_5 的节点频数均不小于 2,故保留所有节点信息。首先从节点 I_3 开始挖掘,矩阵 B 对应节点 I_3 的列为第 3 列,其中元素为 1 的是第 3、5、6、7、8、9 行,保留这些信息得到新的矩阵。

$$B = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, I = \begin{bmatrix} I_1:4 & 0 \\ I_2:5 & 0 \\ I_3:6 & -6 \end{bmatrix}$$

将 I_3 作为一个频繁项集的元素,并得到新的矩阵:

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, I = \begin{bmatrix} I_1:4 & -1 \\ I_2:5 & -5 \end{bmatrix}$$

则 I_2 也作为此频繁项集的元素;再重复一次,可得此频繁项集为 $\{I_1, I_2, I_3\}$,且其所有子集(包括一项和两项)也是频繁项集。

同理从节点 I_4 进行挖掘,其频繁项集为 $\{I_2, I_4\}$;从节点 I_5 进行挖掘,其频繁项集为 $\{I_1, I_2, I_5\}$ 。至此,挖掘结束,所挖掘出来的频繁项集及其所有子集构成了频繁项集的集合。

2 关联规则挖掘算法的比较

FP-Tree 模型的一个关键缺点是维护和更新困难,因为当新数据加入、原数据库更新或者改变支持度阈值时,FP-Tree 算法需要重新扫描数据库来生成 FP-Tree 和条件 FP-Tree。但是 FP-network 模型不存在这个问题,因为 FP-network 是以关联矩阵的形式保存的,而事务-项目关联矩阵中节点的顺序是任意的。例如,若调换节点 I_5 和 I_1 的顺序,可以将上述事务-项目关联矩阵做如下调整:

$$B = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}, I = \begin{bmatrix} I_5:2 & -1 \\ I_2:8 & 0 \\ I_3:6 & -1 \\ I_4:2 & -1 \\ I_1:6 & -6 \end{bmatrix}$$

所产生的 FP-network 如图 3 所示。

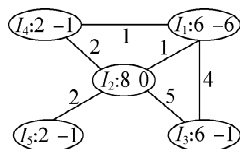


图 3 更新后的 FP-network 示意图

Fig.3 Schematic diagram of updated FP-network

上述 FP-network 的关联规则挖掘首先从 I_3 进行挖掘,得到频繁项目模式为 $\{I_2, I_3, I_1\}$;然后,从节点 I_4 开始挖掘,得到频繁项集为 $\{I_2, I_4\}$;最后从节点 I_1 进行挖掘,得到频繁项目模式为 $\{I_5, I_2, I_1\}$ 。至此挖掘结束,FP-network 不需要重新扫描数据库,只需要对矩阵进行操作,但仍然与上述结果相同。这说明 FP-network 模型与节点的排列顺序无关,克服了 FP-Tree 算法更新维护困难的缺点:

a. 添加新的事务数据时,只需在矩阵 B 增加新的第 j 行和第 i 列,并改变相应的 f_j, n_j ;

b. 改变支持度阈值 $S_{\min}$ 时,无需对矩阵进行任何处理,可直接在原矩阵上进行关联规则挖掘。

综上所述,FP-network 算法与目前主要的关联规则挖掘算法 Apriori、FP-Tree 的比较如表 2 所示。

表 2 关联规则挖掘算法的比较

算法	时间复杂度	空间复杂度	实现情况	更新维护难度
Apriori	生成 1 个频繁 N -项集需要扫描 N 次数据库	存储大量候选集,甚至出现组合爆炸的问题	原理简单	困难
FP-Tree	扫描 2 次数据库	采用递归算法,需要频繁进出栈操作,占用大量内存	树的计算生成十分繁琐	困难
FP-network	仅需扫描 1 次数据库	以布尔矩阵存储,占用内存最小	仅需对矩阵进行处理	方便

3 算例分析

为了进一步验证所提 FP-network 算法的优越性以及展示 FP-network 算法如何应用于智能电网大数据分析,本文采用操作系统为 Windows 10、内存为 4 GB、CPU 为 Intel(R) Core(TM) i5-4430CPU@3.00 GHz 的实验环境,借助 Anaconda 平台,使用 python 语言开发,实现 Apriori、FP-Tree、FP-network 算法的实验测试。

以关联规则挖掘在输电线路故障分析领域的应用为例进行算例分析。采用某省电力公司大数据平台提供的输电线路故障信息,原始信息系统数据量庞大,达到 TB 级,但由于价值密度低,所以对 2010—2017 年线路故障历史数据进行预处理(除噪、清洗、过滤等),得到 1 276 条有效信息,形成线路典型故障事务数据库,其中部分样本如表 3 所示。为满足算法测试需求,复制真实事务信息,可得到包含 12 760 条信息和 127 600 条信息的事务数据库。

对于数据库的“时间”属性,需要进行离散处理,鉴于实际分析需要,可忽略年份信息; T_1 表示春季(3—5月); T_2 表示夏季(6—8月); T_3 表示秋季(9—11月); T_4 表示冬季(12月至次年2月)。其他属性本身就是离散变量,定义了一系列字母变量来代替,预处理后的结果如表 4 所示。

在相同的实验环境下,取 $S_{\min} = 0.5\%$,分别测试 Apriori、FP-Tree、FP-network 算法在不同数据库规模下的运行速率,结果如图 4 所示。

由于图 4 可见,FP-network 算法的运行速率要优于其他 2 种算法,且数据库规模越大,其他 2 种算法,尤其是 Apriori 的运行时间呈指数增长,使得 FP-network 优势更加明显。

表3 线路典型故障事务数据库

Table 3 Transaction database of typical line faults

事务	时间	区域	电压等级/kV	故障元件	故障类型	合闸情况	损失
T_{001}	2011-03-02	北部地区	220	导线及地线	雷击	合闸成功	较轻
T_{002}	2011-11-09	西部地区	110	导线及地线	舞动	合闸失败	较轻
T_{003}	2011-07-06	中部地区	220	导线及地线	大风	合闸失败	较轻
T_{004}	2011-08-17	中部地区	500	绝缘子	外力破坏	合闸失败	一般
T_{005}	2012-04-27	北部地区	220	杆塔及基础	覆冰	合闸成功	一般
T_{006}	2012-06-23	中部地区	500	金具	舞动	合闸成功	严重
T_{007}	2012-02-04	中部地区	220	绝缘子	雷击	合闸成功	一般
T_{008}	2014-06-08	南部地区	220	导线及地线	雷击	合闸成功	较轻
T_{009}	2012-07-31	东部地区	220	绝缘子	外力破坏	合闸失败	严重

表4 预处理后的事务数据表

Table 4 Transaction database after preprocessing

事务	时间	区域	电压等级	故障元件	故障类型	合闸情况	损失
T_{001}	T_1	L_4	V_2	FC_1	FT_3	Y	IN_1
T_{002}	T_3	L_3	V_1	FC_1	FT_2	N	IN_1
T_{003}	T_2	L_5	V_2	FC_1	FT_5	N	IN_1
T_{004}	T_2	L_5	V_3	FC_3	FT_1	N	IN_2
T_{005}	T_1	L_4	V_2	FC_2	FT_4	Y	IN_2
T_{006}	T_2	L_5	V_3	FC_4	FT_2	Y	IN_3
T_{007}	T_4	L_5	V_2	FC_3	FT_3	Y	IN_2
T_{008}	T_2	L_2	V_2	FC_1	FT_3	Y	IN_1
T_{009}	T_2	L_1	V_2	FC_3	FT_1	N	IN_3

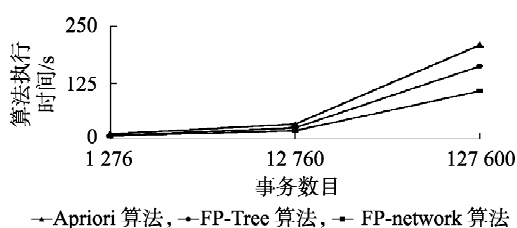


图4 不同数据库规模下的算法执行时间

Fig.4 Executive time of algorithms with different database scales

改变支持度不会改变事务数据库的规模,但是会改变频繁项集的规模。采用包含 127 600 条信息的数据库测试不同支持度下的算法性能,结果如图 5 所示。由图可见:在不同支持度下,FP-network 算法的运行速率均优于另外 2 种算法;调低最小支持度 ($\min_sup$) 后,3 种算法的执行时间都有所增加,FP-network 算法与其他 2 种算法的差距愈加明显。此外,FP-network 算法可以较好地应对支持度的变化,运行速率变化

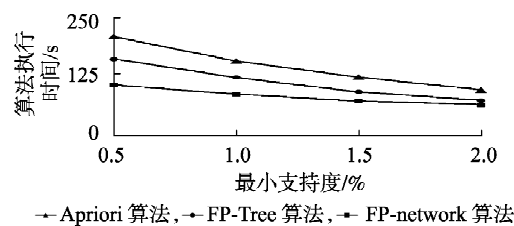


图5 不同最小支持度下的算法执行时间

Fig.5 Executive time of algorithm with different $\min_sup$

幅度较小。

空间复杂度是对算法在运行过程中临时占用存储空间大小的量度,包括程序代码所占用的空间、输入数据所占用的空间和辅助变量所占用的空间。其中,输入数据所占用的空间不随算法的不同而改变;程序代码所占用的空间与算法书写的长短成正比。对于 Apriori、FP-Tree、FP-network 算法而言,存储算法本身的空间仅 kB 级,相较 MB、GB 级的数据库其差异可忽略不计;算法空间复杂度的对比重点在于辅助变量所占用的空间。利用 memory_profiler 模块,监控算法执行过程中所占用的最大内存,结果如图 6 所示。

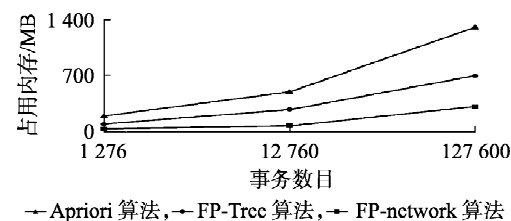


图6 不同数据库规模下算法执行过程中占用的最大内存

Fig.6 Maximum memory in execution of algorithms with different database scales

由图 6 可见,Apriori 算法需要存储大量候选集,FP-Tree 算法需要存储条件树,并需要进出栈操作,占用内存较大,FP-network 算法占用内存空间远小于这 2 种算法,特别是在数据库规模较大时优势更加明显。综上所述,FP-network 算法在处理大型数据库或强实时性问题时性能优势明显,适合电力系统大数据分析。因此,取 $S_{\min} = 0.5\%$,对故障信息数据库进行分析,得到所有频繁项集,进一步挖掘关联规则,结果如图 7 所示,图中连线的粗细表示关联程度的强弱。

如图 7 所示,得到的关联规则有上百条,但是并非所有的关联规则都是有价值的。有些规则关联性十分弱,有些关联性并没有实际的意义。通过计算规则置信度(如式(2)所示),筛选出置信度不低于 75%的规则,得到部分结果如表 5 所示。

$$\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cap Y)}{\text{support}(X)} \quad (2)$$

其中, $\text{support}(\cdot)$ 表示支持度。

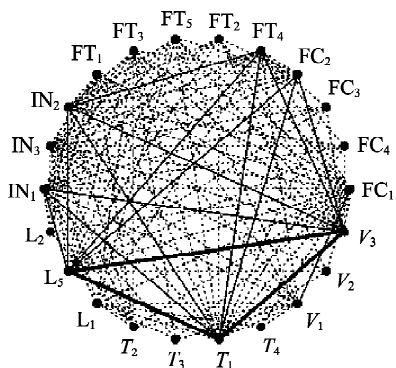


图 7 关联规则挖掘结果图

Fig.7 Results of association rules mining

表 5 关联规则挖掘结果

Table 5 Results of association rules mining

规则	关联规则	支持度/%	置信度/%
1	$\{T_1, L_5, V_2, IN_3\} \rightarrow \{FC_1, FT_2\}$	0.7	100
2	$\{T_2, L_2, FT_1\} \rightarrow \{FC_1\}$	1.4	90
3	$\{V_3, FC_3\} \rightarrow \{T_4, L_5\}$	5.7	85
4	$\{L_1, V_1\} \rightarrow \{FC_1\}$	3.2	91
5	$\{FT_4\} \rightarrow \{T_1, L_5, V_3\}$	6.4	84

结合实际电力知识背景,利用上述挖掘结果可以分析该省的线路故障情况,找到电网中存在的薄弱环节,并针对这些薄弱环节提出改进措施和方案:

a. 从规则 1 可知,该省中部地区 220 kV 线路在 3—5 月份因为导线及地线舞动造成严重影响,因此应做好春季线路舞动预防措施;

b. 从规则 2 可知,该省南部地区在 6—8 月因外力破坏(如违规施工)造成导线及地线故障较多,因此相关部门要采取措施杜绝违规施工;

c. 从规则 3 可知,该省 500 kV 的线路故障大多

是绝缘子故障,主要集中在中部地区的 12 月至次年 2 月,因此检修部门可以在冬季有针对性地对中部地区多加巡查;

d. 从规则 4 可知,该省东部地区的 110 kV 线路故障大多数是导线及地线故障,因此相关部门要优化脆弱地区的线路布局等;

e. 从规则 5 可知,该省线路覆冰故障主要集中在 3—5 月份,并以中部地区的 500 kV 线路居多,因此中部地区应在该月份加强对 500 kV 线路的监测,及时对覆冰进行融化处理。

4 在智能电网中的其他应用

如第 3 节所示,FP-network 可以挖掘多维属性间的关联规则,除输电线路故障分析领域外,其在智能电网中的应用十分广泛。智能电网大数据的来源大致可分为电网外部和电网内部 2 类。电网外部的数据来源包括但不限于互联网信息系统、充电设施管理系统、气象监测系统、地理信息系统(GIS)等;电网内部的数据来源主要包括用电信息采集系统、设备运维管理系统(PMS)、企业资源计划(ERP)系统、SCADA 系统、D5000 系统、95598 客服系统、营销系统等^[7-8]。不同的系统存储着不同的数据信息,同时也对应着不同的业务部门和业务需求。

因此,本文将智能电网中常用的数据和信息系统进行了梳理,建立“数据-系统-业务”体系,如图 8 所示。通过图 8 所示体系,数据的来源和流向一目了然,既可以利用数据来驱动业务发展,也可以基于业务流程和需求进行数据挖掘分析。因此基于智能电网“数据-系统-业务”体系,FP-network 算法在智

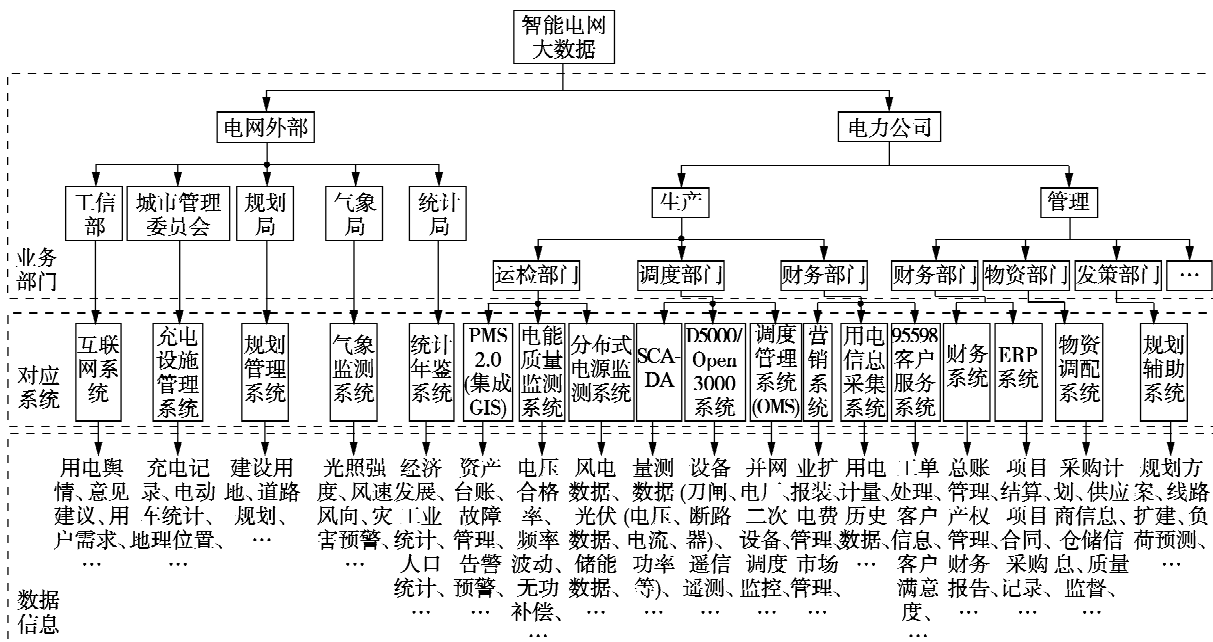


图 8 智能电网“数据-系统-业务”图

Fig.8 Diagram of “Big data-System-Business” in smart grid

能电网中的应用可以根据业务需求分为三大类^[9]:面向用户服务、面向电力公司管理、面向政府决策。本文限于篇幅,着重介绍以下应用^[10-14]。

a. 用户用电行为分析。

配用电环节要做到智能化,需要基于海量用户用电特征数据,如用电类别、时间、客户、行业、电压等级、气象、峰谷负荷等进行分析。通过 FP-network 算法挖掘不同用电行为特征间的关联规则,描述用电行为模式,实现面向用户服务的用电管理、有序用电,面向电力公司的台区负荷预测、用电调度,面向政府的工业发展趋势预测、电价制定等高级应用。

b. 电力系统故障分析。

电力系统的故障数据由用电信息采集系统、PMS、检修运维系统、ERP 系统、GIS、气象信息系统(MIS)等组成,通常包括故障时间、地点、天气、故障类型、故障元件、损失、保护开关动作、恢复时间等一系列属性,通过 FP-network 充分挖掘历史数据多维属性间的关联关系,可以发现故障产生的规律,用于识别电网的薄弱环节、制订检修计划以及故障预警等,避免类似故障的再发生。

c. 电力市场营销策略。

电力公司营销的数据库由售电量、交易电价、用户种类、气象信息、客户关系、客户满意度等组成,应用 FT-network,描述各种影响电量销售的外部因素与售电量、交易电价等之间的关联特征,可以进行需求预测、销售及收入预测,掌握营销业务重点工作,为电力市场营销提供辅助的决策信息。

d. 风光运行优化。

通过对风光发电系统历史运行数据的分析,可得特定工况下气象数据(风速、光照强度、温度等)与机组性能间的关联关系,从而对机组性能(可靠性、经济性、安全性等)进行客观、正确评估,辅助实现风电准确预测、风电场规划、制订出力调度计

划等。

FP-network 在智能电网的应用流程的步骤主要有数据采集、数据预处理、数据挖掘、关联规则分析及结果展示。其中,数据预处理包括噪声清洗、数据离散化处理、缺失值填补等。图 9 以电力系统故障分析为例说明应用的具体流程。

5 结论

鉴于智能电网大数据发展的需要和目前关联规则挖掘算法存在的缺点,本文建立了适合智能电网应用的 FP-network 模型。所得到的主要结论如下:

a. FP-network 模型与 FP-Tree 模型相类似,将关联规则挖掘所需要的数据压缩到一个图上,但是 FP-network 图扩大了存储的事务规模,实际挖掘时需要以矩阵形式进行存储;

b. 与 FP-Tree 模型相同,FP-network 模型同样只能处理分类变量(即离散变量),因此需要预先对事务数据进行离散化;

c. FP-network 模型只需扫描 1 次原数据库,且以矩阵形式存储,尤其对于智能电网大型数据库而言,大幅降低了时间和空间的复杂度;

d. FP-network 模型方便被挖掘数据的更新和维护,因此提高了关联规则挖掘算法的效率;

e. FP-network 模型适合挖掘智能电网大数据,应用范围包括但不限于故障分析、营销策略制订、用电负荷研究、风电运行优化等。

参考文献:

- [1] 彭小圣,邓迪元,程时杰,等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报,2015,35(3):503-511.
PENG Xiaosheng, DENG Diyu, CHENG Shijie, et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE, 2015, 35(3):503-511.
- [2] 宋亚奇,周国亮,朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术,2013,37(4):927-935.
SONG Yaqi, ZHOU Guoliang, ZHU Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology, 2013, 37(4):927-935.
- [3] AGRAWAL R, IMIELSKI T, SWAMI A. Mining association rules between sets of items in large databases[J]. Acm Sigmod Record, 1993, 22(2):207-216.
- [4] HAN T, KAMBER M. Data mining: concepts and techniques[M]. Beijing: Higher Education Press, 2001:143-177.
- [5] SAVASMR A, OMICINSKI E, NAVATHC S. An efficient algorithm for mining association rules[C]//The 21th International Conference on VLDB. Zurich, Switzerland: Morgan Kaufmann Publishers Inc., 1995:432-444.
- [6] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation[C]//The 2000 ACM SIGMOD on Management of Data (SIGMOD 2000). Dallas, Texas, USA: ACM, 2000:1-12.
- [7] TSENG V, SHIE B E, WU C, et al. Efficient algorithms for mining high utility item sets from transactional databases[J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2013, 25

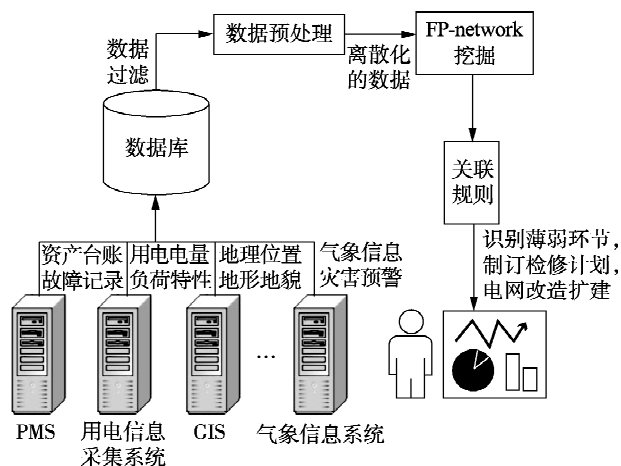


图9 FP-network 在故障分析应用的流程图

Fig.9 Flowchart of FP-network application in fault analysis

- (8):1772-1786.
- [8] 王守相,葛磊蛟,王凯. 智能配电系统的内涵及其关键技术[J]. 电力自动化设备,2016,36(6):1-6.
WANG Shouxiang, GE Leijiao, WANG Kai. Main contents and key technologies of smart distribution system[J]. Electric Power Automation Equipment, 2016, 36(6):1-6.
- [9] 张东霞,苗新,刘丽平,等. 智能电网大数据技术发展研究[J]. 中国电机工程学报,2015,35(1):2-12.
ZHANG Dongxia, MIAO Xin, LIU Liping, et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE, 2015, 35(1):2-12.
- [10] 薛振宇,胡航海,宋毅,等. 基于大数据分析的县公司综合评价策略[J]. 电力自动化设备,2017,37(9):199-204.
XUE Zhenyu, HU Hanghai, SONG Yi, et al. Comprehensive evaluation based on big data analysis for county electric power company[J]. Electric Power Automation Equipment, 2017, 37(9):199-204.
- [11] 郝然,艾芊,肖斐. 基于多元大数据平台的用电行为分析构架研究[J]. 电力自动化设备,2017,37(8):20-27.
HAO Ran, AI Qian, XIAO Fei. Architecture based on multivariate big data platform for analyzing electricity consumption behavior[J]. Electric Power Automation Equipment, 2017, 37(8):20-27.
- [12] 葛磊蛟,王守相,瞿海妮. 智能配用电大数据存储架构设计[J]. 电力自动化设备,2016,36(6):194-202.
GE Leijiao, WANG Shouxiang, ZHAI Haini. Design of storage framework for big data of SPDU[J]. Electric Power Automation Equipment, 2016, 36(6):194-202.
- [13] 徐青山,王文帝,林章岁,等. 面向行业大数据特征挖掘的电力经理指数指标体系的建立与应用[J]. 电力自动化设备,2015,25(7):15-21.
XU Qingshan, WANG Wendi, LIN Zhangsui, et al. Establishment and application of EMI indicator system orienting to massive industrial data mining[J]. Electric Power Automation Equipment, 2015, 25(7):15-21.
- [14] 王德文,孙志伟. 电力用户侧大数据分析与并行负荷预测[J]. 中国电机工程学报,2015,35(3):527-537.
WANG Dewen, SUN Zhiwei. Big data analysis and parallel load forecasting of electric power user side[J]. Proceedings of the CSEE, 2015, 35(3):527-537.

作者简介:



孙丰杰

孙丰杰(1993—),男,山东淄博人,硕士研究生,主要研究方向为智能电网大数据(E-mail:sunfengjie@sjtu.edu.cn);

王承民(1970—),男,辽宁大连人,教授,博士,主要研究方向为电力系统规划与经济运行;

谢宁(1973—),女,新疆乌鲁木齐人,副教授,博士,主要研究方向为电力系统安全稳定分析。

Frequent pattern network model for association rule mining of big data in smart grid

SUN Fengjie, WANG Chengmin, XIE Ning

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Because FP-Tree(Frequent Pattern-Tree) algorithm for association rule mining is hard to achieve and it also has problems in handling database update, a novel FP-network model is proposed, which innovatively compresses the required data into an undirected network and establishes the transaction-item matrix for the data storage and data mining. The proposed FP-network model is suitable for the association rule mining of big data in smart grid. Case study of the application of association rule mining in power transmission line fault analysis is carried out. The results show that the proposed FP-network algorithm not only inherits the merits of the FP-Tree algorithm, but also only needs to scan the database once, which is convenient for database maintenance and update, and improves the efficiency of association rule mining for big data in smart grid.

Key words: smart grid; big data; association rules; FP-Tree algorithm; FP-network algorithm