第 42 卷　第 4 期　　计　算　机　工　程　　2016 年 4 月
Vol. 42　No. 4　　Computer Engineering　　April 2016

·人工智能及识别技术·　　文章编号：1000-3428(2016)04-0190-07　　文献标识码：A　　中图分类号：TP18

基于加权频繁模式树的通信网络告警规则挖掘方法

罗　明[1]，孟传伟[2]，黄海量[1]

（1. 上海财经大学信息管理与工程学院，上海 200433；2. 上海电信科技发展有限公司，上海 200083）

摘　要： 传统通信网络告警处理方法主要由维护专家依据经验判断形成处理规则并固化在网络告警系统中进行实现，然而该人工维护方式难以适应海量数据环境下实时通信告警规则的处理需求。为此，提出一种基于加权频繁模式树（WFP-tree）算法的告警规则自动挖掘方法，将原始告警数据按时间窗口方式进行分段处理，通过 BP 神经网络、支持向量机、层次分析法生成告警设备的权重信息，并采用 WFP-tree 算法自动挖掘加权频繁项集。实验结果表明，与传统 Apriori 和 FP-growth 算法相比，WFP-tree 算法在通信网络告警分析方面具有更好的频繁项压缩效果及更强的重要关联规则发现能力。

关键词： 通信网络告警；关联规则；权重因子；加权频繁项集；FP-growth 算法；加权频繁模式树算法；支持度

中文引用格式： 罗　明，孟传伟，黄海量. 基于加权频繁模式树的通信网络告警规则挖掘方法[J]. 计算机工程，2016，42(4)：190-196.

英文引用格式： Luo Ming，Meng Chuanwei，Huang Hailiang. Alarm Rule Mining Method in Telecommunication Network Based on Weighted Frequent Pattern-tree[J]. Computer Engineering，2016，42(4)：190-196.

Alarm Rule Mining Method in Telecommunication Network Based on Weighted Frequent Pattern-tree

LUO Ming[1]，MENG Chuanwei[2]，HUANG Hailiang[1]

（1. College of Information Management and Engineering，Shanghai University of Finance and Economic，Shanghai 200433，China；
2. Shanghai Telecom Science & Technology Development Co.，Ltd.，Shanghai 200083，China）

【Abstract】 Traditional communication network alarm correlation rules are often manually done by experts and coded into network fault management systems. However，the artificial maintenance method is difficult to meet the huge amounts of data processing requirements of real-time communication alarm rules. To solve this problem，this paper proposes an automatic alarm rule mining method based on Weighted Frequent Pattern-tree（WFP-tree）algorithm. It uses the sliding window method to convert raw data into alarm transactions，and employs BP neural network，Support Vector Machine（SVM）and Analytic Hierarchy Process（AHP）methods to generate the weight information of alarm equipment. Finally，it uses WFP-tree algorithm to automatically generate the weighted frequent itemset. The experimental results show that，the WFP-tree algorithm performs better in frequent itemset compression and important domain correlation rule finding compared with Apriori and FP-growth algorithms.

【Key words】 communication network alarm；correlation rule；weighted factor；weighted frequent itemset；FP-growth algorithm；Weighted Frequent Pattern-tree（WFP-tree）algorithm；support degree

DOI： 10.3969/j.issn.1000-3428.2016.04.034

1　概述

通信网络管理的主要工作之一是对网络的日常运行进行实时监控，以确保通信网络能够高效、可靠、低成本地运行。但一方面由于通信网络存在网络节点规模大、网络复杂度高、网络异构性强的特点，导致网络中任何一个节点产生的故障都有可能会影响、波及到全网中其他节点，从而产生大量的冗

基金项目： 上海市科技创新行动计划基金资助项目（13511505200）；上海市科技人才计划基金资助项目（14XD1421000）；上海财经大学 2014 年研究生创新基金资助项目（CXJJ-2014-438）。

作者简介： 罗　明（1974－），男，高级工程师、博士研究生，主研方向为数据挖掘、模式识别；孟传伟，高级工程师、博士；黄海量（通讯作者），教授、博士。

收稿日期： 2015-03-12　　**修回日期：** 2015-04-10　　**E-mail：** luoming@189.cn

余告警。另一方面在同一个通信网络中由底层到高层又垂直分布着动力环境网、传输网、数据网、交换网、无线网等多种异构网络[1],并且同一种网络中又运行着不同设备厂商、不同型号的设备,这些情况使网络障碍管理成为通信网络管理工作中的难点。尤其是近年来通信网络 IP 化、扁平化的发展趋势要求在全省级别甚至跨省级别上统一实现对全网的调度和集约化运维,这些为如何提升网络障碍管理工作带来了更高的挑战。为在海量原始通信设备告警数据中实现过滤、收敛、分类和处理关键性故障类告警信息,必须采用相关性分析的手段实现对告警信息的关联识别。为此,本文结合通信网络的实际特点,设计一组衡量网元权重的权重因子,分别采用 BP 神经网络、支持向量机(Support Vector Machine,SVM)和层次分析法(Analytic Hierarchy Process,AHP)对网元权重进行分类和评价,在获得权重信息的基础上,提出一种 FP-growth 的加权频繁模式树(Weighted Frequent Pattern-tree,WFP-tree)算法。

2　相关研究

近年来,国内外学术界对通信网络告警相关性分析处理方法的研究大致可以分为以下 3 类:

(1)采用手工配置规则的方法[2]。目前国内外电信运营商普遍采用该方法。基本思路是维护专家结合专业知识和网络拓扑来逐条确定对原始告警信息的相关性处理规则,并将其手工维护到告警处理系统中。该方法的优点是规则生成过程简单、易快速部署实施,对原始告警信息易实现按告警类型的快速收敛,适用于网络结构相对简单,不经常调整的小规模网络。该方法的缺点是由于处理规则是人工逐条生成,因此只能做到按告警类型进行较粗粒度的告警定位,且告警系统本身不具备自学习和自调整的能力,一旦网络结构发生经常性或较大调整时(如网络层次布局发生变化、核心层网元设备位置调整和更换、新增监控专业、新增监控子网),采取手工配置规则库的方法需要花费大量的人力对现有规则库进行重新梳理和调整,同时由于不同地域设备情况的差异和专业分工的不同等客观因素的存在,很难通过这种方法实现对跨专业领域设备和跨地域告警的相关性分析工作。

(2)采用专家经验知识库的方法[3-4]。该方法利用专家对过去发生的一些告警事件的处理事例来帮助解决新的问题。该方法的优点是有一定的自学习能力,过去被成功解决的案例可以积累下来为将来处理类似告警提供参考。该方法的一个显著性缺点是由于专家的处理案例仅限于某一专业领域或某个局部地域情况下的特殊案例,因此,一方面不能适用于大规模、动态、复杂的通信网络发展要求;另一方面系统判断处理告警的过程较复杂,难以适应海量数据下的实时性通信告警处理需求。

(3)采用数据挖掘的方法[5]。该方法近年来随着数据挖掘理论的发展,已逐渐获得了学者们的重视,并且已经成为通信网络告警相关性研究的主流方法。该方法的优点是利用各种成熟的数据挖掘理论和技术,能够方便地实现告警规则的自动生成和动态更新,能够适应当今通信网络结构复杂、动态变化的要求。该方法的缺点是模型的设计和实施难度较大,对挖掘结果需要进一步做推广化处理。

近年来国内外关于采用数据挖掘方法进行告警相关性分析的研究主要有:文献[6]提出基于关联规则方法的 Aprori 挖掘算法来寻找告警频繁项集,实现告警处理规则的自动生成。当 FP-growth 算法[7]在取得成功后,基于 FP-growth 算法来解决通信网络告警频繁项挖掘的研究也越来越多[5,8],由于采用 FP-growth 算法的优点在于挖掘过程中不需要生成大量的候选集,并且对全库数据的遍历也是一次完成,因此大大降低了系统内存开销,提高了挖掘速度,适用于针对通信复杂网络条件下的海量数据挖掘。此外,文献[9]提出一种基于滑动窗口的 Top-k 概率频繁项集增量查询算法,以实现对关联规则中的频繁项集的增量更新发现。文献[10]提出一种基于模式挖掘和聚类分析的自适应关联告警模型。文献[11]采用贝叶斯网络方法建立通信网络相关性模型,采用 EM 算法对隐变量进行学习,并进行网络告警相关性分析。文献[12]提出一种采用动态模糊关联规则挖掘算法生成光网络设备的告警关联规则。

由于不同层次、不同专业类型的通信网元在全网中的重要程度不同,因此在挖掘过程中不能简单地将网元看成平等关系,而应针对不同的网元设备和告警类型赋予不同的权重值,这样才能够挖掘出更重要的信息,并过滤掉一些维护人员不关心的信息。现有关联规则挖掘算法大都是基于 Aprori 算法和 FP-growth 算法,并不适合于加权模式的挖掘,而一些基于 FP-growth 加权模式的研究[13-14]实用性不强,并且不能直接用于针对通信网络的告警挖掘工作,因此,有必要开发一种新的方法来解决通信行业的关联规则挖掘问题。

3　告警权重设计

本文首先选择适合通信行业特点的权重因子,其次采用 BP 神经网络、SVM 和 AHP 这 3 种方法分别预测生成告警权重,最后采用测试集数据对 3 种

方法的预测效果进行评价。

3.1 权重因子选择

影响网络设备告警权重大小的因素通常有以下5个方面：

(1)网络层次。一般来说通信网络的级别分为4级，其重要性大小依次为：主干层 > 核心层 > 汇聚层 > 接入层。

(2)告警类型。告警类型是指该告警是否紧急的重要标志，一般情况下通信网络设备的告警类型分为：事件，警告，次要，重要，紧急。

(3)专业类别。一般情况下网络设备的专业类别大致可分为：传输，交换，动力环境，无线和数据。

(4)出入度。网络节点的出入度(特别是出度)不同，对网络的影响也不同。但对于一个复杂庞大的通信网络，一般很难完全获取网络上所有设备的出入度信息，因此，依靠出入度指标来评价权重的方法并不可行。

(5)告警时间。网络设备在不同告警时间发出的告警重要程度不同。但告警时间的紧急程度可以通过告警类型的方式来体现，因此，在处理过程中可不考虑该方面的因素。

综上所述，本文选取了网络层次、告警类型、专业类型3个因素作为本文的权重评价因子，并设计了权重因子的取值范围，如表1所示。

表1 权重因子及其取值范围

权重因子	取值范围
专业类型	1为传输、2为动环、3为数据、4为无线
网络层次	1为接入层、2为汇聚层、3为核心层、4为主干层
告警类型	1为事件、2为警告、3为次要、4为重要、5为紧急

3.2 权重生成

本文采用的权重值生成方法为：(1)有监督学习方式(BP神经网络分类法和SVM分类法)；(2)无监督学习方式(AHP层次分类法)。第(1)类方法中分类变量为设备告警权重，所对应的权重值为离散值，分别为0，0.1，0.2，0.3。第(2)类方法中的权重取值范围是(0，1)的连续数值。下面分别予以说明：

(1)用BP神经网络分类法生成权重，设计步骤具体如下：

1)确定输入层的神经元和连接权。输入层的神经元个数 $N=3$，分别为网络层次、告警类型和专业类型，输入样本的向量矩阵可表示为一个 $m \times n$ 的矩阵：

$$\boldsymbol{TR}^0 = \begin{bmatrix} tr_{11} & tr_{12} & tr_{13} \\ tr_{21} & tr_{22} & tr_{31} \\ \vdots & \vdots & \vdots \\ tr_{m1} & tr_{m2} & tr_{m3} \end{bmatrix} = [\xi_1 \quad \xi_2 \quad \xi_3]$$

其中，m 为输入记录数(训练集数或测试集数)。

2)确定连接权和隐藏层、输出层阈值。本文中隐藏层的神经元个数为 M 个(隐藏层神经元数目无特殊要求，本文 M 分别取3，10，20)，输出层的神经元(预测权重类别)为4个。

3)确定期望输出向量。期望输出向量 $\boldsymbol{y}=(y_1, y_2, y_3, y_4)=(0, 0.1, 0.2, 0.3)$ 为模型训练集数据中的专家实际打分的设备告警权重值。

4)确定隐藏层和输出层传输函数。采用sigmoid函数作为隐藏层和输出层的传递函数，即：

$$f(x) = g(x) = \frac{1}{1-\mathrm{e}^{-x}},\ -\infty < x < \infty$$

5)采用BP算法反复迭代计算预测输出向量，并且不断调整连接权值，当输出层神经元的输出 C_t^k 与期望输出误差：$|e^k(k)| = \left| \sum_{t=1}^{L} (y_t - C_t^k(k)) \right| < \varepsilon$，$\forall \varepsilon > 0$ 时，迭代过程结束，输出预测结果。

(2)用SVM支持向量机分类法生成权重，设计步骤具体如下：

1)确定训练集输入向量和分类向量。训练集输入向量集合：$\boldsymbol{TR}=(\boldsymbol{tr}_1^{\mathrm{T}}, \boldsymbol{tr}_2^{\mathrm{T}}, \cdots, \boldsymbol{tr}_m^{\mathrm{T}})$，其中，$m$ 为数据集记录数；$\boldsymbol{tr}_i = \begin{bmatrix} tr_{i1} \\ tr_{i2} \\ tr_{i3} \end{bmatrix}$，$tr_{i1}$，$tr_{i2}$，$tr_{i3}$ 分别表示训练集第 i 条记录的设备类型、网络层次、告警类型数据值。训练集分类向量：$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$，$\boldsymbol{y}_i$ 表示对应训练集第 i 条记录的4维分类向量(行向量)，其取值范围为[1 0 0 0]，[0 1 0 0]，[0 0 1 0]，[0 0 0 1]，分别与专家打分的设备告警权重值(0，0.1，0.2，0.3)相对应。

2)分别采用线性分类器(LinearSVC)、线性核分类器(SVC with linear kernel)、径向基函数核分类器(SVC with RBF kernel)、多项式核分类器(SVC with polynomial kernel)4种分类器对权重数据进行分类。

3)比较上述方法的测试结果，选择预测准确率最高的模型对全体数据进行预测。

(3)用AHP层次分类法生成权重，本文采用的层次模型如图1所示。

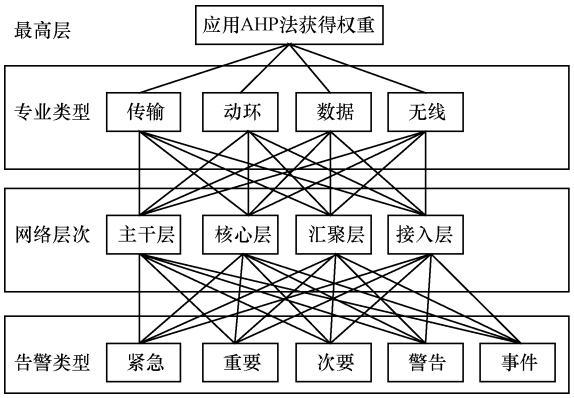

图 1　AHP 层次模型

设计步骤具体如下:

1)采用 1 ~ 9 标度法[15]设计调查问卷,获得专家对各准则层项目的两两评价意见。

2)构造层次判断矩阵。专业类型准则层 1 个,网络层次准则层 4 个,告警类型准则层 4 个,共计 9 个判断矩阵。

3)计算每个判断矩阵的最大特征根和特征向量,对每层进行层次单排序和一致性检查。

4)计算层次总排序及一致性检查。本文所需要确定的最终权重项共计为:$C_4^1 \times C_4^1 \times C_5^1 = 90$ 个。

5)以最终权重矩阵 $\boldsymbol{W}$ 为判定标准,为所有的网元设备赋权重。其中,$\boldsymbol{W} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & w_1 \\ a_{21} & a_{22} & a_{23} & w_2 \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & w_m \end{bmatrix}$,$a_{i1}, a_{i2}, a_{i3}, w_i$ 分别为第 i 项专业类型、网络层次、告警类型组合及其权重值。

4　WFP-tree 算法设计

4.1　算法定义

定义 1(事务数据库)　设 $I = \{i_1, i_2, \cdots, i_n\}$ 是一个由告警项目组成的集合(itemset),数据库 $D = \{t_1, t_2, \cdots, t_n\}$ 是由一系列具有唯一 TID 的设备告警事务构成,每个事务 $t_i (i = 1, 2, \cdots, n)$ 都是由对应于告警项目集合 I 的一个子集 X 构成($X \subset I$,X 中的元素 i_t 是唯一不重复的),则 D 称为告警事务数据库。

定义 2　对于给定的告警项目集 $I = \{i_1, i_2, \cdots, i_n\}$,集合 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ 称为项目集 I 的设备告警权重集合。为了消除不同项集尺寸大小对权重的影响,项集 X 的权重用项集 X 中各项权重和的平均值来表示,记作 $\frac{1}{k}\left(\sum_{i_j \in X} \omega_j\right)$。

定义 3　一个项集 X 的加权支持度表示为:$wsupport(X) = \frac{1}{k}\left(\sum_{i_j \in X} \omega_j\right) \times support(X)$,当 $wsupport(X) \geqslant wminsupport$ 称该项集为加权频繁项集。

定义 4　在 k-项集(L_k)条件下,包含项集 X 的最大期望权重值表示为:$\omega(X, k) = \frac{1}{k}\left(\sum_{i_j \in X, j=1}^{q} \omega_j + \max\left(\sum_{j=1}^{k-q} \omega_{r_j}\right)\right)$,其中,项集 X 的权重和表示为:$\sum_{i_j \in X, j=1}^{q} \omega_j$;项集 $L_k - X$(在集合 L_k 中 X 的补集)的最大权重和表示为:$\max\left(\sum_{j=1}^{k-q} \omega_{r_j}\right)$;包含 X 的 k-项集最小加权支持度计数表示为:$B(X, k) = \frac{wminsup \times T}{\omega(X, k)}$,其中,$T = |D|$ 为 D 中的所有事务记录数。

定义 5　对一个项集 $X = \{I_1, I_2, \cdots, I_j\}$,令其最小权值项为 ω_{X_i},若存在项集 $Y = X \cup Z$ $(X, Z \notin \varnothing)$,当 Z 项集中包含的所有项的权重都小于 ω_i 时,则 Y 项集称为 X 项集的低阶超集,X 项集称为 Y 项集的高阶子集。

定义 6(加权潜在频繁模式)　当尺寸为 p 的项集 X 的支持度计数 $supportcount(X)$ 满足大于最小支持度计数 $B(X, k)$ $(p < k \leqslant L, L$ 为项集最大尺寸)的条件时,项集 X 称为加权潜在频繁 p 模式。

为解决加权频繁项集单调向下不闭合性质[16]条件下的剪枝问题,容易证明加权频繁项集存在如下一些性质,本文将其用于 WFP-tree 算法的频繁项集处理策略中,使 WFP-tree 算法在继承了 FP-growth 算法的占用内存空间少、执行效率高等优点的基础上,又解决了 FP-growth 不适用于加权频繁项挖掘的问题。

性质 1　对于高阶子集 X 和低阶超集 $Y(Y = X \cup Z(X, Z \notin \varnothing))$之间存在如下不等式性质:$\frac{1}{k}\sum_{i=1}^{k} \omega_{Z_i} < \frac{1}{k+j}\sum_{i=1}^{k+j} \omega_{Y_i} < \frac{1}{j}\sum_{i=1}^{j} \omega_{X_i}$。

证明:令 Z 中项目的最大权重值为 β,则有:

$$\frac{1}{k}\sum_{i=1}^{k} \omega_{Z_i} \leqslant \frac{1}{k}(k \times \beta) \Rightarrow \frac{1}{k}\sum_{i=1}^{k} \omega_{Z_i} \leqslant \beta$$

令 X 中项目的最小权重值为 α,则根据定义 4 可知 X 中各项目的权重之和至少大于等于 $j \times \alpha$,由此可得 $\frac{1}{j}\sum_{i=1}^{j} \omega_{X_i} \geqslant \frac{1}{j}(j \times \alpha) \Rightarrow \frac{1}{j}\sum_{i=1}^{j} \omega_{X_i} \geqslant \alpha$。因为 $Y = X \cup Z$,且 $\alpha > \beta$,所以 $\frac{1}{k+j}\sum_{i=1}^{k+j} \omega_{Y_i} > \frac{1}{k+j}(k+j) \times \beta$

$=\beta, \frac{1}{k+j}\sum_{i=1}^{k+j}\omega_{Y_i} < \frac{1}{k+j}(k\times\alpha + j\times\alpha) = \alpha$，从而得到 $\frac{1}{k}\sum_{i=1}^{k}\omega_{Z_i} < \frac{1}{k+j}\sum_{i=1}^{k+j}\omega_{Y_i} < \frac{1}{j}\sum_{i=1}^{j}\omega_{X_i}$。

性质 2 如果一个项集 Y 为加权频繁项集，则 Y 的任意一个高阶子集 X 也必然是加权频繁项集（证明从略）。

性质 3 一个尺寸为 $(k+1)$ 的加权频繁项集 Y 必定存在一个尺寸为 k 的子集 X 为加权频繁项集。

性质 4 当一个尺寸为 p 的项集 X 不满足定义 6的加权频繁模式条件时，则该项集及其所有超集都不是加权频繁项集，应从树中剪除。

4.2 算法描述

WFP-tree 算法具体如下：

输入 告警事务数据库 T，最小加权支持度阈值 *wminsup*，加权列表 *weightlist*

输出 加权频繁项集

变量说明：L_1-*list* 为 1-模式项集；PL_1-*list* 为加权潜在 1-模式频繁项集；*maxsize* 为最大项集尺寸；*masterTree* 为初始的 FP-tree；$B(X_i,k)$ 为项集 X_i 的 k 项集最小支持度计数；$\omega(X_i,k)$ 为项集 X_i 的 k 项集最大期望权重；$subtree_{x_i}$ 为项目 x_i 的条件子树；*wfrequentlist* 为加权频繁项集

```
1.  WFP-tree(T,wminsup,weightlist)
2.  L1-list = ∅, PL1-list = ∅,
    maxsize = 0, weightedfrequentlist = ∅;
3.  L1-list, supportcount, maxsize = scan(T);//扫描 T,
    //获得 L1-list 和列表项目的 supportcount, maxsize
4.  sort(L1-list);//对 L1-list 中的项目按 supportcount
    //值降序排序
5.  L1-list.item.weight = weightlist;//根据 weightlist 表
    //将相应的权重值赋给 L1-list 中项目
6.  for each xi ∈ L1-list do
7.    {For(k = 1;k≤maxsize;k++)
8.      ω(xi,k) = 1/k(ωxi + max(Σ_{j=1}^{k-1} ωxj));
9.      B(xi,k) = (wminsup × |T|)/ω(X,k);}
    //根据定义 4 分别计算 ω(xi,k)和 B(xi,k)的值
10.     If supportcount(xi) ≥ B(xi,k) then
11.       {PL1-list = PL1-list ∪ xi; break;}
    //根据性质 2、性质 3 进行潜在频繁项集判断，并将 xi
    //加入潜在频繁项集列表 PL1-list 中
12. cleanTrans = Clean(PL1-list,T);
    //删除非潜在加权频繁项
13. masterTree = Build(cleanTrans,PL1-list);
    //采用文献[7]的方法构建 FP-tree 主树
14. wfrequentlist = wfequentlist + PL1-list;
15. for each xi in masterTree.HeadTable do
16.     {subtreexi = FP-tree(xi);
        //采用文献[7]的方法生成 xi 的条件子树
17. wfrequentlist = wfequentlist + Search(xi);}
        //对条件子树subtreexi 执行频繁项集查找操作
18. 输出结果集：wfrequentlist
```

WFP-tree 算法中的新增主要子程序描述如下：

（1）*Clean*(PL_1-*list*, T)

输入 PL_1-*list* 为加权潜在 1 项集列表，T 为所有事务记录

输出 删除非潜在加权项以及按支持度排序的事务记录集

```
1.  For each trans in T do{
2.    For eachxi in X do{
3.      If xi in PL1-list then;continue;//保留 xi 项;
4.      Else:trans = trans - xi;//将 xi 项从 trans 事务中
      //删除;
5.    Sort(xi);//对 trans 事务中的项 xi 按支持度计
      //数大小降序排列;
6.  }}
7.  返回 T
```

（2）*Search*(x_i)

输入 x_i 为条件子树 $subtree_{x_i}$ 的叶节点 x_i

输出 查找加权频繁项集 *wfrequentlist*

```
1.  k = length(xi) + 1, treehigh = length(subtreexi);
2.  ω(xi,k) = 1/k (ωxi + max(Σ_{j=1}^{k-1} ωxj));
3.  B(xi,k) = (wminsup × |T|)/ω(xi,k)//计算 xi 项的最小
    //加权支持度计数;
4.  If supportcount(xi) > B(xi,k) then
    {xi 分别与子树中的前缀项组成长度为 k+1 的项集 β;
5.    wfrequentlist = wfrequentlist + β
6.    For each βi in β
7.      {xi = βi;
8.       If (length(xi) < treehigh):执行 Search(xi)}}
9.  Else{//根据性质 4 将不满足条件的项集剪除
10. xi 与子树中的前缀项构成长度为 k+1 的组合 β;
11.   For each βi in β
12.     {xi = βi;
13.      If (length(xi) < treehigh):执行 Search(xi)}}
14. 返回 wfrequentlist
```

5 实验结果与分析

本文从权重因子生成方法的选择、加权和非加

权挖掘算法执行效果比较 2 个方面进行实验验证。本文方法均采用 python 编程语言在 python3.2 编译环境下实现。实验环境:CPU 为 Intel 酷睿 i7(2.1 GHz 4 核,2 级缓存为 1 GB,3 级缓存为6 MB);内存为 8 Gb;硬盘为 1 TB;操作系统为 Windows8.0;数据库为 mysql 5.0。

5.1　实验数据设置

本文采用的实验数据全部为某电信公司生产数据,为该公司 2 个规模较大的地市级分公司 2014 年 9 月 1 日 –2014 年 9 月 15 日期间的传输、数据、动环、无线 4 个专业所产生的全部原始告警数据,共计 1 049 966条,提炼生成的唯一告警项(item)为 56 571个。原始告警数据分布为如表 2 所示。

表 2　原始告警数据分布

专业类型	告警记录数量	网元数量	告警类型数量
传输	549 824	24 064	158
数据	356 962	15 189	33
动环	93 020	2 672	363
无线	50 160	1 375	81

5.2　不同权重生成方法对频繁项集挖掘的影响

BP,SVM,AHP 挖掘效果如表 3 所示。其中,*wminsup* 为最小加权支持度;*minconf* 为最小置信度;由滑动窗口为 10 s,步长为 5 s 的方法生成实验所用告警事务数据,共计 240 212 条;重要率指挖掘所生成的关联规则被专家认定为重要规则的比率,对一条规则而言:重要为 1,不重要为 0。

表 3　BP,SVM,AHP 方法挖掘效果比较

方法	*wminsup* = 0.84% *minconf* = 90%		*wminsup* = 1.68% *minconf* = 90%		*wminsup* = 2.52% *minconf* = 90%		*wminsup* = 3.36% *minconf* = 95%		*wminsup* = 1% *minconf* = 90%		*wminsup* = 0.1% *minconf* = 90%	
	频繁项集数	关联规则数(重要率/%)	频繁项集数	关联规则数(重要率/%)	频繁项集数	关联规则数(重要率/%)	频繁项集数	关联规则数(重要率/%)	频繁项集数	关联规则数(重要率/%)	频繁项集数	关联规则数(重要率/%)
BP	28	2(100)	13	1(100)	5	0	1	0	25	1(100)	459	30(96)
SVM	48	3(100)	16	0	9	0	4	0	40	2(100)	690	28(100)
AHP	0	0	0	0	0	0	0	0	0	0	3	0

重要规则是必须关注处理的。从表 3、图 2、图 3的数据可以看出,由于 AHP 法所产生的单项权重过小,导致挖掘的结果最差,甚至将 *wminsup* 值设为0.05% 时(即最小支持度计数为 120 条)所挖掘的频繁项集也只有 14 条,因此该方法不适用于通信告警挖掘应用。

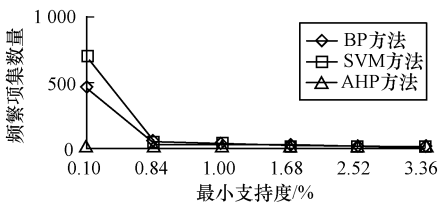

图 2　BP,SVM,AHP 方法的频繁项集挖掘趋势

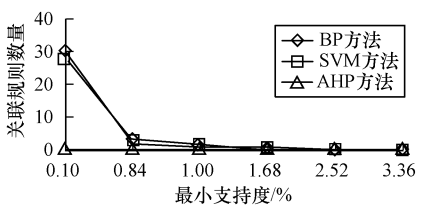

图 3　BP,SVM,AHP 方法的关联规则生成趋势

从 BP 和 SVM 方法的挖掘效果来看,当最小支持度阈值由 0.1% ~0.84% 变化时,2 种方法产生的频繁项集及关联规则都呈线性下降趋势,而最小支持度阈值大于 0.84% 后所产生的频繁项集数目较少。因此,在实际工作中要特别注意加权最小支持度阈值的设置问题。SVM 在同等条件下所生成的频繁项集个数要稍多,且规则重要性略高于 BP 方法。这说明 SVM 方法更适合通信告警挖掘应用。

5.3　FP-growth 与 WFP-tree 算法挖掘效果比较

在不同最小支持度阈值的情况下,观察 FP-growth 算法和 WFP-tree 算法(采用 SVM 生成的权重)所生成的频繁项集、关联规则数量以及规则重要率。在最小支持度阈值、最小置信度阈值设置为固定值(90%)的情况下,采用 FP-growth 算法和 WFP-tree 算法对 240 212 条事务数据进行挖掘并分析挖掘结果。

从表 4、图 4、图 5 可以看出,在不考虑权重影响的情况下,FP-growth 算法所挖掘的频繁集数量要远高于 WFP-tree 算法(特别是当最小加权支持度低于 1% 的情况),同时生成的关联规则数量也远高于 WFP-tree 算法所生成的规则数量。这是因为 FP-growth 算法产生的频繁项集数据中包含了大量出现频次高但非重要的告警信息。这些信息的存在一方面导致所挖掘关联规则重要率降低,特别是当最小支持度小于 0.1% 之后,规则重要率指标急剧恶化;另一方面使很多用户关心的重要设备和紧急障碍的告警信息不能被挖掘,FP-growth 算法产生的大量关

联规则也不利于维护和告警障碍的判断。采用 WFP-tree 算法可以避免以上问题,其重要率指标不随最小支持度的降低而恶化。

表 4 FP-growth 与 WFP-tree 算法挖掘效果比较

最小支持度/%	FP-growth 算法			WFP-tree 算法		
	频繁项集数	关联规则数	重要率/%	频繁项集数	关联规则数	重要率/%
0.10	8 300	258	76.74	690	28	100.00
0.84	452	37	100.00	48	3	100.00
1.00	340	24	100.00	40	2	100.00
1.68	164	13	100.00	16	0	
2.52	103	7	100.00	9	0	
3.36	72	4	100.00	4	0	

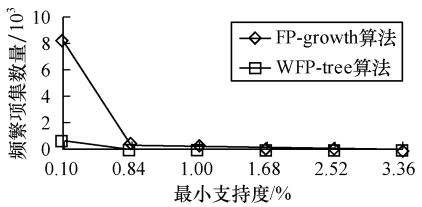

图 4 FP-growth 与 WFP-tree 算法的频繁项集挖掘趋势

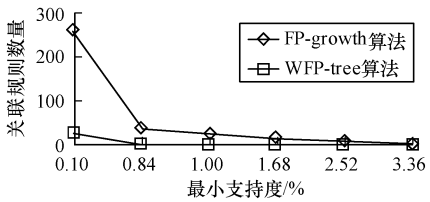

图 5 FP-growth 与 WFP-tree 算法的关联规则生成趋势

6 结束语

本文将告警权重与 FP-growth 算法相结合,提出一种面向通信网络告警相关性分析的加权频繁模式树挖掘算法 WFP-tree,通过一系列实验验证了该算法的有效性。在今后研究中,将在 WFP-tree 算法的基础上,开展告警因果性分析、重大故障预测、告警拓扑生成等方面的研究。

参考文献

[1] 韦乐平.电信技术发展的趋势和挑战[J].重庆邮电大学学报:自然科学版,2010,22(5):545-550.

[2] Dattatraya V K, Shaila D A. A Universal Object Oriented Expert System Frame Work for Fault Diagnosis[J]. International Journal of Intelligence Science,2012,2(3):63-70.

[3] 王 伟,芦东昕,唐 英.基于专家系统的网络故障管理系统的设计[J].计算机工程与设计,2005,26(11):3031-3033.

[4] Raquel B,Pedro L,Volker W,et al. Knowledge Acquisition for Diagnosis Model in Wireless Networks[J]. Expert Systems with Applications,2009,36(3):4745-4752.

[5] 李彤岩.基于数据挖掘的通信网络相关性分析研究[D].成都:电子科技大学,2010.

[6] Klemettinen M,Mannila H,Toivonen H. Rule Discovery in Telecommunication Alarm Data[J]. Journal of Network and Systems Management,1999,7(4):395-423.

[7] Jia Weihan,Jian Pei,Yi Wenyin,et al. Mining Frequent Patterns without Candidate Generation: A Frequent-pattern Tree Approach[J]. Data Mining and Knowledge Discovery,2004,8(1):53-87.

[8] Liu Ting. The New Algorithms of Weighted Association Rules Based on Apriori and FP-growth Methods[J]. TELKOMNIKA Indonesian Journal of Electrical Engineering,2014,12(5):4071-4078.

[9] 王 爽,王国仁.基于滑动窗口的 Top-K 概率频繁项查询算法研究[J].计算机研究与发展,2012,49(10):2189-2197.

[10] 田志宏,张永铮,张伟哲,等.基于模式挖掘和聚类分析的自适应告警关联[J].计算机研究与发展,2009,46(8):1304-1315.

[11] 邓 欹,孟洛明.基于贝叶斯学习的告警相关性分析[J].计算机工程,2007,33(12):40-42.

[12] 吴 简,李兴明.基于动态模糊关联规则推理的光网络故障管理[J].光电工程,2012,39(7):13-25.

[13] Unil Y. Efficient Mining of Weighted Interesting Patterns with a Strong Weight and/or Support Affinity[J]. Information Sciences,2007,177(17):3477-3499.

[14] Vidya V. Mining Weighted Association Rule using FP-tree[J]. International Journal on Computer Science and Engineering,2013,5(8):741-752.

[15] Saaty T L. The Analytic Hierarchy Process[M]. New York, USA:McGraw-Hill,1980.

[16] Cai C H, Fu A W C, Cheng Cheng, et al. Mining Association Rules with Weighted Itemsets[C]//Proceedings of 1998 International Symposium on Database Engineering & Applications. Washington D. C., USA: IEEE Press, 1998:68-79.

编辑 陆燕菲

(上接第 189 页)

[10] Zangwill W I. Convergence Conditions for Nonlinear Programming Algorithms[J]. Management Science, 1969,16(1):1-13.

[11] Luenberger D G, Ye Y. Linear and Nonlinear Programming[M]. Berlin,Germany:Springer,2008.

[12] Bache K, Lichman M. UCI Machine Learning Repository[EB/OL]. [2014-12-05]. http://archive.ics.uci.edu/ml.

[13] Liu Yi, Jin Rong, Jain A K. Boostcluster: Boosting Clustering by Pairwise Constraints[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:ACM Press,2007:450-459.

[14] Jing Liping,Ng M K,Huang J Z. An Entropy Weighting k-means Algorithm for Subspace Clustering of High-dimensional Sparse Data[J]. IEEE Transactions on Knowledge and Data Engineering,2007,19(8):1026-1041.

[15] Deng Zhaohong,Choi K S,Chung F L,et al. Enhanced Soft Subspace Clustering Integrating Within-cluster and Between-cluster Information[J]. Pattern Recognition, 2010,43(3):767-781.

编辑 陆燕菲