

基于加权关联规则的护理人员职场偏差行为研究

王冬燕, 余嘉元

(南京师范大学 心理学系, 江苏 南京 210097)

摘 要: 利用加权关联规则方法提取护理人员职场偏差行为问卷研究中的规则, 样本包括 539 名护理人员。鉴于偏差行为影响因素重要性的不一致, 基于传统的关联规则 Apriori 算法, 设计并实现了加权关联规则的挖掘算法, 应用于护理人员偏差行为的研究。实验表明, 与 Apriori 算法相比, 加权关联规则方法提取出更多的规则, 并且可以增强某些规则的支持度和置信度, 使这些规则更容易被发现, 说明该方法是可行的。

关键词: Apriori 算法; 加权关联规则; 职场偏差行为; 护理

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2012)03-0117-04

On the Research of Nurses' Workplace Deviance Behavior Based on Weighted Association Rules

WANG Dong-yan, YU Jia-yuan

(Department of Psychology, Nanjing Normal University, Nanjing 210097, China)

Abstract: Weighted association rules is applied to extract the rules in the questionnaire of the nurses' workplace deviant behavior. The sample includes 539 nurses. According to the importance of factors that influence workplace deviant behavior is inconsistent, based on the traditional Apriori algorithm for association rules, designing and implementing the mining algorithm of weighted association rules, to applied to the research of nurses' workplace deviant behavior. It was shown that, compared with the Apriori algorithm, weighted association rules can extract more rules, and enhance the support and confidence of some of the rules, to make the rules easier to find. It shows that the method is feasible.

Key words: Apriori algorithm; weighted association rules; workplace deviant behavior; nurses

1 引言

随着护理人员工作负荷的普遍加重, 关注其身心健康, 确保护理质量尤为关键。目前, 护理行业普遍存在员工犬儒主义现象(对组织/主管怀抱敌意并作出负面内在归因)。犬儒主义可能导致职场偏差行为发生^[1], 损害护理品质; 若护理人员具备正向幽默风格, 可能减低职场偏差行为发生。因此, 挖掘职场偏差行为与其影响因素之间的规则, 可以有效的监控导致护理人员职场偏差行为发生的因素, 及时矫正, 从而提高护理质量。

关联规则挖掘是数据挖掘领域一个重要的研究课题。Agrawal 于 1993 年提出布尔型关联规则问

题, 之后提出了经典的 Apriori 和 Apriori TID 算法^[2]。文献[3]中详细介绍了关联规则挖掘的各种算法。然而关联规则挖掘是针对各个因素相互独立且具有同等重要性的情况而进行研究的, 但在现实生活中, 尤其在心理学邻域, 每个影响因素是有轻重之分的, 关联规则分析得到的结果仅是粗略估计。在这种情况下, 使用加权关联规则进行挖掘, 可以获取更有价值的信息。文献[4]提出布尔型属性加权关联规则的概念。文献[5-7]则介绍了一些改进的加权关联规则的算法。

本文中, 影响职场偏差行为的因素较多, 且影响程度不一致, 用传统的关联规则方法, 不容易挖掘出小概率规则或潜在的规则, 本文提出采用基于

Apriori 算法的加权关联规则算法来挖掘与职场偏差行为有关的规则。

2 加权关联规则

关联规则定义: 设 $I = \{i_1, i_2, \dots, i_m\}$ 是项的集合, 记 D 为事务 T 的集合, 这里事务 T 是项的集合, 并且 $T \subseteq I$. 对应每一个事务有唯一的标识, 记作 TID. 设 X 是一个 I 中项的集合, 如果 $X \subseteq I$, 那么称事务 T 包含 X . 一个关联规则是形如 $X \Rightarrow Y$ 的蕴涵式, 这里 $X \subset I, Y \subset I$, 并且 $X \cap Y = \emptyset$. 规则 $X \Rightarrow Y$ 在 D 中的支持度 (support) 是事务集中包含 X 和 Y 的事务数与所有事务数之比, 记为 $\text{support}(X \Rightarrow Y)$, 即 $\text{support}(X \Rightarrow Y) = |\{T: X \cup Y \subseteq T, T \in D\}| / |D|$. 规则 $X \Rightarrow Y$ 在事务集中的置信度 (confidence) 是指包含 X 和 Y 的事务数与包含 X 的事务数之比, 记为 $\text{confidence}(X \Rightarrow Y)$, 即 $\text{confidence}(X \Rightarrow Y) = |\{T: X \cup Y \subseteq T, T \in D\}| / |\{T: X \subseteq T, T \in D\}|$. 给定一个事务集 D , 挖掘关联规则就是产生支持度和置信度分别大于用户给定的最小支持度和最小置信度的关联规则。

2.1 加权关联规则概念

在上述关联规则概念的基础上, 文献[8]是这样定义加权关联规则的: 给定项目集合 $I = \{i_1, i_2, \dots, i_n\}$, 为表征项目的重要性, 为每一个项目 i_j 赋以权重 w_j , 其中 $w_j \in [0, 1], j = \{1, 2, \dots, n\}$.

定义 1 关联规则形如 $X \Rightarrow Y$ 的加权支持度 (weighted support) 为

$$\left[\sum_{i_j \in X \cup Y} w_j \right] (\text{support}(X \cup Y))$$

定义 2 称 k -项目集为频繁项目集, 若其加权支持度不低于最低加权支持阈值 w_{minsup} , 即

$$\left[\sum_{i_j \in X \cup Y} w_j \right] (\text{support}(X \cup Y)) \geq w_{\text{minsup}}$$

定义 3 若 $X \cup Y$ 是频繁项目集, 则其置信度不低于最低置信阈值 minconf .

2.2 加权关联规则算法

传统 Apriori 算法思想分两步: 首先找出所有支持度不小于用户指定的最小支持度 minsup 约束的所有属性组合, 即频繁项集, 记为 L_k , 然后利用频繁项集 L_k 产生满足最小置信度 minconf 约束的所有规则。

关联规则两个非常重要的性质: 频繁项集的子集是频繁项集, 非频繁项集的超集是非频繁项集。本

文算法与 Apriori 相比, 主要有两点改动: (1) 频繁项的子集不一定是频繁项, 因此在计算频繁项的时候必须遍历所有子集, 不能基于 1 项频繁集。 (2) 在支持度的计算时, 需要在出现频率的基础上, 乘以权重之和。由于权重的引入可能会导致置信度大于 1, 而这不符合概率意义, 鉴于它的实际意义, 本文把置信度超过 1 的全部视为 1。

基于 Apriori 算法的加权关联规则算法如下:

```
(1) for(k=1; k<5; k++) // 获取 1-5 项频繁集
(2) {
(3)   for each transaction t ∈ D
(4)     { // 扫描集合 D 获取出现频率
              Ct = subset(D, k); // 检索所有子集
(5)       Ct.count++;
(6)     }
(7)   Lk = {sup(Ct) ≥ min_sup}
(8) }
(9) for (k=2; k<5; k++) // 计算规则
(10) {
(11)   for each transaction t ∈ Lk
(12)   {
(13)     c = subset(Lk); // 检索所有子集
(14)     if (sup(t)/sup(c) ≥ min_conf)
(15)       outputRule(c ⇒ (t - c));
(16)   }
(17) }
```

由于需要检索所有子集, 问卷的题目和答案的取值都有范围, 且每个事务都一致, 故使用数组来存储题号和答案, 数组的值表示出现频率, 提高运行效率。

3 数据准备

3.1 数据属性

护理人员职场偏差行为影响因素较多, 本研究主要采用问卷法从心理契约违反 (CSK)、犬儒主义 (WRA)、差序气氛 (DA)、幽默风格 (HS)、离职意图 (TL)、人口统计学 (DG) 六个因素进行考察, 以及偏差行为 (DB) 本身的表现。问卷共包括七个部分, 由台湾佛光大学赖惠德教授 (2010) 编制提供。其中, 幽默风格又可具体细分为四种类型: 产生连击型 (HAF)、自我提升型 (HSE)、攻击型 (HAG) 和自我贬抑型 (HSD); 人口统计学又分为性别 (SEX)、年龄 (AGE)、教育程度 (EDU) 等八项; 偏差行为本身分为四种行为表现: 员工偏差行为_人际偏差 (PDB_P)、员工偏差行为_组织偏差 (PDB_O)、组织公民行为_利同事 (OCB_P) 和组织公民行为_利组织 (OCB

—O)。本研究中是对划分到最后一个等级的因素属性进行挖掘分析,提取影响职场偏差行为的规则。问卷采用李克特6级量表(Likert Scale)形式,陈述具体情况,被测试者根据自身实际作答。每组陈述有“非常不同意”、“不同意”、“有点不同意”、“有点同意”、“同意”和“非常同意”六种回答,标记为1、2、3、4、5、6。

3.2 属性权重设定

本文利用SPSS18.0进行主成分分析法来确定各属性权值。主成分分析法是将原来众多具有一定相关性的指标(如 P 个指标),重新组合成一组新的互相无关的综合指标来代替原有指标。通常数学上的处理就是将原来 P 个指标作线性组合,作为新的综合指标。最经典的做法就是用 F_1 (选取的第一个线性组合,即第一个综合指标)的方差来表达,即 $\text{var}(F_1)$ 越大,表示 F_1 包含的信息越多。因此在所有线性组合中选取的 F_1 应该是方差最大的,故称 F_1 为第一主成分。如果第一主成分不足以代表原来 P 个指标的信息,再考虑选取 F_2 即第二个线性组合,为了有效地反映原来信息, F_1 已有的信息就不需要再出现在 F_2 中,用数学语言表达就是要求 $\text{cov}(F_1, F_2)=0$,则称 F_2 为第二主成分,依此类推可以构造出第三、第四、…、第 P 个主成分。数学模型为:

$$F_1 = a_{11}X_1 + a_{21}X_2 + \dots + a_{p1}X_p$$

$$F_2 = a_{12}X_1 + a_{22}X_2 + \dots + a_{p2}X_p$$

...

$$F_p = a_{1m}X_1 + a_{2m}X_2 + \dots + a_{pm}X_p$$

其中 $a_{1i}, a_{2i}, \dots, a_{pi}, (i=1, 2, \dots, m)$ 为 X 的协方差矩阵 $\sum$ 的特征值 λ 对应的特征向量, $X_1, X_2, \dots, X_p$ 是原始变量经过标准化处理的值。 $A = (a_{ij})_{p \times m} = (a_1, a_2, \dots, a_m)$, $R_{ai} = \lambda_i a_i$, R 为相关系数矩阵, λ_i, a_i 是相应的特征值和单位特征向量 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 。最后根据各主成分中每个指标前的系数、其特征值、在主成分中的贡献率算出每个指标的权重^[9]。

问卷共收集到651个样本,剔除不合格问卷及极端数值问卷(z 分数在3个 δ 之外),共有有效样本539个。规则提取时,选取重点研究的八项属性:心理契约违反、犬儒主义、差序气氛、离职意图、幽默风格(两类)和员工偏差行为(两类)。数据表形式如表1所示。

4 算法实现及结论

4.1 加权关联规则提取

本文采用C++语言实现Apriori算法和加权关联规则算法。取最小支持度为0.2,最小置信度为0.8,得出的加权关联规则共146条,选择几例列举如表2所示。

综合分析表2,可得出以下结论:(1)从规则1中可以看出护理人员认同单位对其心理契约的良好履行,则较不容易产生攻击型幽默风格。(2)规则2显示护理人员有犬儒主义倾向的,不易在工作中使

表1 539组测量数据

属性 (权值)	CSK (0.26)	WRA (0.24)	DA (0.24)	TL (0.38)	HAG (0.62)	HSD (0.6)	PDB_P (0.34)	PDB_O (0.26)
1	4	5	4	3	3	2	1	1
2	3	4	4	4	3	1	2	2
...	...	...	...	...	...	...	...	...
539	5	6	3	4	3	2	4	3

表2 护理人员职场偏差行为加权关联规则

规则数	加权关联规则	支持度	置信度
1	CSK=4 $\Rightarrow$ HAG=3	0.20	1.00
2	WRA=4 $\Rightarrow$ HAG=3 HSD=3	0.20	1.00
3	DA=4 $\Rightarrow$ PDB_P=2 PDB_O=2	0.26	1.00
4	DA=4 HAG=3 HSD=3 $\Rightarrow$ PDB_P=2	0.29	0.85
5	HAG=3 TL=3 $\Rightarrow$ PDB_P=2 PDB_O=2	0.23	0.96
6	HAG=3 HSD=3 $\Rightarrow$ PDB_P=2 PDB_O=2	0.34	0.81
7	HSD=2 $\Rightarrow$ PDB_P=1 PDB_O=1	0.21	0.87

用攻击型和自我贬抑型幽默风格。(3)从规则 3、4、6、7 分别可以看出当护理人员认同单位差序格局存在时,员工越少使用攻击型和自我贬抑型幽默风格,职场偏差行为的产生也越少。(4)规则 5 表明护理人员离职意图不明显的,也很少会产生职场偏差行为。为了保证护理质量,必须有效防止护理人员偏差行为的产生。

因此从上述结论可以给予管理者启示,例如通过心理契约的良好履行,来加强护理人员正确的职业态度;或者多关注护理人员幽默风格的表现,来判断其偏差行为发生的可能性,从而及时予以沟通矫正。

4.2 加权关联规则算法与 Apriori 算法比较

用 Apriori 算法和加权关联规则算法对表 1 数据分别进行实验,比较分析结果见表 3。

表 3 Apriori 算法和加权关联规则算法比较

实验次数	最小支持度	最小置信度	Apriori 算法	加权关联规则算法
1	0.1	0.8	65	789
2	0.1	0.9	19	639
3	0.1	1.0	2	539
4	0.2	0.8	9	146
5	0.2	0.9	1	125
6	0.2	1.0	0	109

由表 3 不难看出,对于同一最小支持度和最小置信度得到的关联规则数,加权关联规则数总是大于传统关联规则数。可以看出,加权关联规则可以挖掘出比传统关联规则更多的潜藏规则。例如:当选用最小支持度为 0.1,最小置信度为 1.0,得到的传统关联规则数为 2,得到的加权关联规则数为 539,且其中规则 $DA=5 \text{ HSD}=2 \Rightarrow PDB_P=1 \text{ PDB_O}=1$ 出现在加权关联规则中而不在传统关联规则中。它在传统关联规则中的支持度为 0.10,置信度为 0.89,而在加权关联规则中的支持度为 0.14,置信度为 1.0。因此使用加权关联规则方法可以增强了某些规则的支持度和置信度,使这些规则更容易被发现。用于护理人员偏差行为的研究中,可以发掘出更多潜在的影响因素,及时纠正,以便更好的提高护理人员的护理质量。

5 结束语

本文探讨了基于加权关联规则的护理人员职场

偏差行为的分析方法。实验研究表明,采用加权关联规则提取职场偏差行为相关影响因素之间的规则是可行的。但是,该研究一方面还需要质量更高的样本数据的支持,另一方面也需要对加权关联规则算法的优化,使其能更好的避免冗余规则的产生。由于心理研究的特殊性,加权关联规则的挖掘方法应该会成为心理学领域数据挖掘的有效方法之一。

参考文献:

- [1] 黄怡祯. 组织中的雇佣关系与职场偏差行为[D]. 台北: 中原大学心理学系, 2006.
- [2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C] // Proceedings of the 1993 ACM SIGMAD interational Conference on Management of data. NY, USA: ACM, 1993: 207-216.
- [3] 佟强, 周元春, 阎保平. 关联规则挖掘算法[J]. 微电子学与计算机, 2005, 22(6): 68-72.
- [4] Cai C H, Fu A W C, Cheng C H, et al. Mining association rules with weighted items[C] // Proceedings of the Database Engineering and Application Symposium Cardiff, UK: IEEE, 1998: 68-77.
- [5] 张智军, 方颖, 许云涛. 基于 Apriori 算法的水平加权关联规则挖掘[J]. 计算机工程与应用, 2003, 39(14): 197-199.
- [6] 尹群, 王丽珍, 田启明. 一种基于概率的加权关联规则挖掘算法[J]. 计算机应用, 2005, 25(4): 805-807.
- [7] 李成军, 杨天奇. 一种改进的加权关联规则挖掘算法[J]. 计算机工程, 2010, 36(7): 55-57.
- [8] 欧阳为民, 郑诚, 蔡庆生. 数据库中加权关联规则的发现[J]. 软件学报, 2001, 12(4): 612-619.
- [9] 张文霖. 主成分分析在满意度权重确定中的应用[J]. 市场研究, 2006(6): 18-22.

作者简介:



王冬燕 女, (1984—), 博士研究生. 研究方向为认知心理学。



余嘉元 男, (1949—), 博士生导师. 研究方向为认知心理学。