关联规则挖掘技术的研究与应用

王晓宇,秦 锋,程泽凯,邹洪侠
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

摘 要:数据挖掘技术是目前广泛研究的数据库技术。它可以从大量的数据中提炼出有用的潜在的信息,人们可利用这些信息改进工作,提高效率。介绍了数据挖掘中关联规则挖掘算法的基本概念、思想、步骤及当前的一些算法,并在此基础上改进 Apriori 算法,研究分析了某市儿童意外伤害数据的主要特点,以及造成儿童意外伤害的主要危险因素。在对挖掘的结果进行分析后,提出预防意外伤害的措施。描述了关联规则挖掘在此类调查中的应用、发展及存在的问题。

关键词:数据挖掘;关联规则;意外伤害调查;数据
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)05-0220-04

Investigation and Application of Association Rules Mining

WANG Xiao-yu, QIN Feng, CHENG Ze-kai, ZOU Hong-xia
(Dept. of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: Data mining is the recent extensive study of database technology. It can extract the potential and useful informations from large amounts of data. People can use these informations to improve works and efficiency. Introduces the basic concepts, ideas, steps and some of the current algorithms of association rules mining. Based on the improvement of the algorithm of Apriori, analyze the main risk factors of the children's accident injury. Mining in the analysis of the results, point the measures to prevent the suddenness. At the same time described application of association rules mining in such investigations, development and the existing problems.

Key words: data mining; association rules; research of accident injury; data

0 引 言

关联规则挖掘(Association Rules Mining)是为了在数据库中发现关联关系,它是数据挖掘(Data mining, DM)最先研究的问题之一,也是数据挖掘的一个主要研究方向[1]。关联规则可直观地表达数据中项集(变量的各种取值)间的联系。这种联系并不是基于某种特定的分布,依靠数据在特定模型中的多次迭代拟和而来,而是根据项集在数据资料中出现的概率来构建。因而,这种方法有异于传统的统计学方法,其优势在于结果明确,容易解释。在实际应用中,关联规则即是从大量的、不完全的、有噪声的、模糊的、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。它提取出的信息将有助于人们把握和预测行业发展规律,从而更好地制定发展计划和规避风险[2]。而当变量类型比较复杂,变量取值的分布不定并难于转换,或者各变量不独立,不能满足传统统计学方法的要求时,通过关联规则的挖掘,可以得到数据中隐含于变量取值中的信息。

收稿日期:2008-08-28
基金项目:安徽省自然科学研究重点项目(KJ2007A051);安徽省自然科学研究项目(2006KJ061B)
作者简介:王晓宇(1984-),男,安徽巢湖人,硕士研究生,研究方向为人工智能、数据挖掘、机器学习;秦 锋,教授,硕士生导师,研究方向为人工智能、计算机软件理论和网络应用技术。

1 关联规则挖掘技术

1.1 关联规则挖掘技术的基本原理

关联规则挖掘是由 R. Agrawal 等人提出来的,关联规则是描述数据库中数据项之间某种潜在关系的规则[3]。它的基本概念为:设 $I=\{i_1,i_2,\cdots,i_m\}$ 为数据项集合,设 D 为与任务相关的数据集合,也就是一个交易数据库,其中的每个交易 T 是一个数据项子集,即 $T\subseteq I$;每个交易均包含一个识别编号 TID。设 A 为一个数据项集合,当且仅当 $A\subseteq T$ 时就称交易 T 包含 A。一个关联规则就是具有"$A\Rightarrow B$"形式的蕴含式;其中有 $A\subset I$,$B\subset I$ 且 $A\cap B=\varnothing$。规则 $A\Rightarrow B$ 在交易数据集 D 中成立,具有支持度 s,其中 s 是 D 中交易包含 $A\cup B$(即 A 和 B 二者)的百分比,这是概率 $P(A\cup B)$。如果 D 中包含 A 的事务同时也包含 B 的百分比是 c,则规则 $A\Rightarrow B$ 在交易数据集 D 中具有置信度 c。这

是条件概率 $P(B \mid A)$。即 Support$(A \Rightarrow B) = P(A \cup B)$，Confidence$(A \Rightarrow B) = P(B \mid A)$。满足最小支持度阈值和最小置信度阈值的关联规则就称为强规则，即关联规则[3]。这两个阈值均在 0% 到 100% 之间[4]。

1.2 关联规则分类

可以从多个角度对关联规则进行分类[5]：

(1)基于规则中处理的变量类型，关联规则可以分为布尔型和数值型。布尔型关联规则处理的数据都是离散的、分类化的，它显示了这些变量间的关系。数值型关联规则处理的变量包含有数量信息(数值型变量)，它表示的是属性值之间的关联关系。在挖掘数值型关联规则时要先对数值型变量进行离散化处理，再对处理后的数据进行挖掘。

(2)基于规则中的抽象层次，可以分为单层关联规则和多层关联规则。单层关联规则中，所有变量都不考虑现实中多个层次的区分，多层关联规则就能够体现实际生活中概念的层次性。

(3)基于规则中涉及到的数据的维数，可分为单维关联规则和多维关联规则。单维关联规则只涉及数据表中的单个维(字段)间的关系，多维关联规则表示的是多个维之间的关系。根据是否允许同一字段在规则中重复出现，多维关联规则又可以分为维间关联规则和混合关联规则。

(4)通过对关联规则进行一定的约束和限制，可以生成某些具有针对性的特殊类型的关联规则，这样得到的规则通常是实际工作中人们最感兴趣，也是对实践活动最具指导意义的规则。

1.3 关联规则挖掘算法

关联规则的挖掘一般可分成两个步骤：(1)发现所有的频繁项集，根据定义，这些项集的支持度至少应等于(预先设置的)最小支持度阈值；(2)根据所获得的频繁项集，产生相应的强关联规则。根据定义这些规则必须满足最小支持度阈值和最小置信度阈值。第一步的工作是相当费时的，而第二步在第一步的基础上很容易实现，因此关联规则挖掘算法的性能主要由第一步决定[6]。

Apriori算法是关联规则挖掘的基本算法。针对单维、单层、布尔型关联规则挖掘的 Apriori 算法[7,8]的核心是候选项集的生成。它基于以下性质(Apriori 性质)[9]：如果一个项集是频繁项集，那么它的所有子集都是频繁项集。反之，如果一个项集的某个子集不是频繁项集，那么这个项集也不是频繁项集。算法的第一步是找出所有频繁项集；第二步再由频繁项集产生强关联规则。首先产生一阶频繁项集 L_1，然后是二阶频繁项集 L_2，直到有某一阶的频繁项集 L_R 为空，这时算法停止。这里在第 k 次循环中，过程先产生候选 k - 项集的集合 C_k，C_k 中的每一个项集是对两个只有一个项不同的属于 L_{k-1} 的频繁项集做一个$(k-2)$ - 连接来产生的。C_k 中的项集是用来产生频繁项集的候选集，最终的频繁项集 L_k 必须是 C_k 的一个子集。最后再由前述定义判断那些频繁项集的关联规则。

1.4 当前 Apriori 算法的改进

目前国内外专家学者针对不同关联规则类型的挖掘和数据含量不同的关系型数据库提出多种关联规则挖掘算法，并在减少数据读取和内存占用的目标下对 Apriori 算法进行了改进。主要有：采用杂凑技术改进候选集生成过程的 DHP(Direct Hashing and Pruning)算法；采用分而治之的思想来解决内存不足问题的分块挖掘算法(Partition)；抽样算法(Sampling)；动态项集计数算法 DIC(Dynamic Itemset Counting)[10,11]等。

2 关联规则在儿童意外伤害中的应用

意外伤害是21世纪儿科和儿童保健领域的前沿课题。既往学者认为，意外伤害是意料不到的事件，没有一定的形式或可预见性，因而是不可避免的。随着医学的发展，现在普遍认为，意外伤害虽然是突发事件，但存在内部的发展规律[12]。意外伤害是受伤者 - 动因 - 环境诸多因素综合作用的结果，分析儿童伤害的危险因素，可更好地提出伤害的预防策略和措施。数据挖掘(Data Mining)作为一种新兴的数据库技术，为从海量数据中萃取有用的、隐含的规律提供了捷径。鉴于针对儿童意外伤害调查的数据庞杂和信息处理的动态要求，数据挖掘技术为分析其中潜在的规律和对日后做出相应的预防提供了可能。

2.1 关联规则挖掘的基础数据

为了研究儿童意外伤害是否具有潜在的规律，本课题从某市疾病控制中心提取100个案例在关系表中进行关联规则挖掘。把其整合到关系表中进行关联规则挖掘。

表1为整合之后的信息。

表 1　原始数据表

编号	年龄	性别	家庭状况	受伤程度	受伤原因	责任承担
001	14 岁	男孩	一般	轻伤	跌倒	个人
002	10 岁	女孩	较好	重伤	交通事故	社会
003	6 岁	女孩	较好	重伤	动物咬伤	社会
004	13 岁	男孩	较差	轻伤	烧烫伤	个人
005	10 岁	男孩	一般	重伤	跌倒	社会
……	……	……	……	……	……	……
096	16 岁	男孩	较好	轻伤	交通事故	个人
097	1 岁	男孩	较好	重伤	交通事故	社会
098	15 岁	女孩	较差	轻伤	烧烫伤	个人
099	五个月	女孩	较好	重伤	动物咬伤	社会
100	14 岁	男孩	较差	轻伤	动物咬伤	个人

2.2 基于概化的数据预处理

为了更好地进行关联规则挖掘，要对表 1 中的基础信息进行基于概化的数据预处理(见表 2)，具体的概化处理方法为：

① 用符号 A 描述年龄，把年龄进行分段概化为：A1(≤4 岁)，A2(≤8 岁)，A3(≤12 岁)，A4(≤ 16 岁)。

② 用符号 B 描述性别，B1 表示“女”，B2 表示“男”。

③ 用符号 C 描述家庭状况，C1 表示较好，C2 表示一般，C3 表示较差。

④ 用符号 D 表示受伤程度，轻度受伤为 D1，重度受伤的表示为 D2。

⑤ 用符号 E 表示受伤原因，分别概化为：E1(跌倒)，E2(动物咬伤)，E3(交通事故)，E4(烧烫伤)。

⑥ 用符号 F 表示责任承担，F1 表示由个人承担责任，F2 表示由社会承担责任。

表 2 对表 1 中的数据进行概化的结果

编号	年龄	性别	家庭状况	受伤程度	受伤原因	责任承担
001	A4	B2	C2	D1	E1	F1
002	A3	B1	C1	D2	E3	F2
003	A2	B1	C1	D2	E2	F2
004	A4	B2	C2	D1	E4	F1
005	A3	B2	C2	D2	E1	F2
.....	.....	.....	.....	.....	.....	.....
096	A4	B2	C1	D1	E3	F1
097	A1	B2	C1	D2	E3	F2
098	A4	B1	C2	D1	E4	F1
099	A1	B1	C1	D2	E2	F2
100	A4	B2	C3	D1	E2	F1

2.3 关联规则挖掘过程

由关联规则的概念和表 2 的概化结果，可得出项目集合为{A1，A2，A3，A4，B1，B2，C1，C2，C3，D1，D2，E1，E2，E3，E4，E5，F1，F2}，表 3 即为事件处理信息表。其目的是要从事件信息表中分析几种常见的儿童意外伤害内在的关联规则。假设关联规则的支持度至少为 40%，置信度至少为 80%。进行关联规则挖掘过程。

表 3 事件信息表

TID	项 ID 的列表
T1	A4 B2 C2 D1 E1 F1
T2	A3 B1 C1 D2 E3 F2
T3	A2 B1 C1 D2 E2 F2
T4	A4 B2 C2 D1 E4 F1
T5	A3 B2 C2 D2 E1 F2
...	....
T96	A4 B2 C1 D1 E3 F1
T97	A1 B2 C1 D2 E3 F2
T98	A4 B1 C2 D1 E4 F1
T99	A1 B1 C1 D2 E2 F2
T100	A4 B2 C3 D1 E2 F1

(1)首先利用基于事物压缩的 Apriori 算法找出频繁项集如表 4 所示。

表 4 第一大项目集

频繁 1_ 项集 L_1 (支持度计数)
A1(50)
B1(40)
B2(60)
C1(50)
D1(50)
D2(50)
E1(50)
F2(50)

(2) 改进 Apriori 算法。

由于在儿童意外伤害调查中关注的是能够造成各种后果的诱导因素，且面对的数据量较为庞大，改进的 Apriori 算法在保证提取规则的数量的同时，也增加了条件判断的数量。对初始数据进行概化后，读入数据集 D 时假设最小支持度，生成频繁 1_ 项集 L_1，当 $k=2$ 时，通过 $(k-1)$ 频繁集项目集生成 $k-$ 候选集 C_k，然后通过 C_k 生成 $k-$ 频繁集项目集 L_k，直到 L_k 为空时，合并 L_1-L_k 得到频繁集 L 和最大频项目集 L_M，生成关联规则并输出强关联规则。

具体过程如图 1 所示。

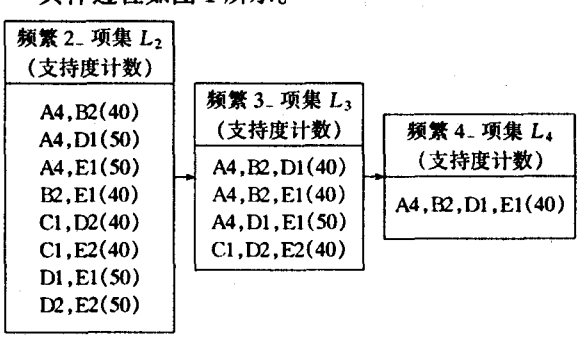

图 1 生成频繁集

(3)得出的结论。

找出支持度至少为 40% 而且置信度至少为 80% 的强关联规则。由以上算法得出的其有关而且实用的强关联规则为：(A4，B2，D1)→E1 (置信度为 100%，支持度为 40%)。此规则可解释为几种常见的儿童意外伤害中，跌倒是儿童意外伤害的首要原因，并且，是发生率最高的非致死性伤害。男童的发生率是女童的三倍。10～14 岁的儿童发生率最高。

2.4 关联规则挖掘结果的指导作用

因意外伤害造成儿童受伤在某市住院治疗的 0～

3 岁年龄组小儿发生意外伤害的主要危险因素是坠落 47.6%、跌伤 23.3%、车祸 23.53%,与某市疾病控制中心所提供的研究结论一致[13]。3～6 岁、6～16 岁两个年龄组发生意外伤害的最主要因素均为车祸,分别占 80.77%、77.8%,坠落伤两组分别占 15.4%、2.14%,扎、砸伤两组分别占 3.85%、7.41%。车祸于文献报导明显增高,分析原因为我国机动车增多。砸伤、坠落伤两年龄组较文献报导明显降低,分析原因与家庭、学校、社会高度重视、加强管理、强化自身安全教育有关。

3　结束语

利用关联规则挖掘方法分析出了隐藏在儿童意外伤害的海量数据背后的有效信息,进而从中获取有意义的部分,并从中挖掘出内在规律,以达到防患未然、防微杜渐的目的。然而关联规则挖掘技术在针对儿童意外伤害中的应用不只是文中提到的这几个方面,还包括如何针对这几种常见的意外形式做出有效的保障和预防工作。数据挖掘技术是具有广阔前景的数据处理与分析技术,它将在有大量信息的数据处理工作中发挥不可估量的作用。

参考文献:

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. 范　明,孟小峰,等译. 北京:机械工业出版社,2001.

[2] 李　锋,孙根年. 旅游目的地灾害事件的影响机理研究[J]. 灾害学,2007,22(3):134－137.

[3] 陈　安,陈　宁. 数据挖掘技术及应用[M]. 北京:科学出版社,2006:50－110.

[4] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proceedings of the ACM SIGMOD International Conference Management of Date. Washington: [s.n.],1993:207－216.

[5] Konias S, Chouvarda I, Vlahavas I, et al. A novel approach for incremental uncertainty rule generation from databases with missing values handling: application to dynamic medical databases[J]. Med Inform Internet Med,2005,30(3):211－215.

[6] 秦亮曦,史忠植. 关联规则研究综述[J]. 广西大学学报:自然科学版,2005(4):48－55.

[7] Agrawal R, Skirant R. Fast Algorithms for Mining Association Rules in Large Databases[J]. California: IBM Almaden Research Center,1994.

[8] 谈恒贵,王文杰,李克双. 频繁项集挖掘算法综述[J]. 计算机仿真,2005(11):1－4.

[9] 梁　循. 数据挖掘:建模、算法、应用和系统[J]. 计算机技术与发展,2006,16(1):1－4.

[10] Wolff R, Schuster A. Association rule mining in peer－to－peer systems[J]. IEEE Trans Syst Man Cybern B Cybern, 2004,34(6):24－26.

[11] 姚卫新,黄丽华. 智能数据分析在医学领域的应用综述[J]. 计算机工程,2004(7):3－5.

[12] 霍　焱,张永文,张莉琴. 儿童意外伤害的健康教育研究[J]. 护理研究,2003(19):130－131.

[13] 尚　磊,徐勇勇,江　逊. 儿童意外伤害住院的情况调查[J]. 疾病控制杂志,2001(3):233－235.

(上接第 219 页)

的设置等,采用不同编码方式和不同大小的滑动窗口,使用 BP 神经网络分类器对蛋白质二级结构进行预测,得到了较好的结果,尤其是 Profile 编码方式获得了较好的实验结果,证明了实验方法的可行性和有效性。为进一步提高预测的精确度,可采用串联 BP 神经网络或其它多种神经网络结合的方式,调整隐含层神经元个数,采用富含生物进化信息的其它编码方式作为输入层数据等方法进行改进,这也是今后需要做的工作。

参考文献:

[1] 马栋萃. 基于改进 BP 神经网络预测蛋白质二级结构[D]. 北京:北京工业大学,2005.

[2] Wang Zhixin. The current situation and prospect of protein structure prediction[J]. Chemistry of Life, 1998,18(6):19－22.

[3] 郑婷婷,毛军军. Profile 覆盖算法在蛋白质二级结构预测中的应用[J]. 计算机技术与发展,2007,17(9):171－172.

[4] 王勇献. 蛋白质二级结构预测的模型与方法研究[D]. 长沙:国防科学技术大学,2004.

[5] 葛哲学,孙志强. 神经网络理论与 MATLAE7 实现[M]. 北京:电子工业出版社,2007.

[6] McGinnis S, Madden T L. BLAST: at the core of a powerful and diverse set of sequence analysis tools[J]. Nucleic Acids Res,2004,32:20－25.

[7] Fariselli P, Pazos F, Valencia A, et al. Predication of protein－protein interaction sites in heterocomplexes with neural networks[J]. Eur. J. Biochem, 2002,269:1356－1361.

[8] Kabsch W, Sander C. Dictionary of Portein seconday structure: patten Recognition of hydrogen－bonded and geometrical features[J]. Biopolymers,1983,22:2577－2637.

[9] 阮晓钢,孙海军. 编码方式对蛋白质二级结构预测精确度的影响[J]. 北京工业大学学报,2005,31(3):229－231.