基于信息扩张机制的意外规则处理新方法研究

胡 健[1],罗家国[1],杨炳儒[2]
(1.江西理工大学,江西 赣州 341000;
2.北京科技大学,北京 100083)

摘 要:文中利用知识发现领域中的信息扩张机制,研究提出一种对意外规则取舍和理解的新处理方法——三度(规则支持度、可信度、充分性因子)变化趋势分析,再结合规则前件、后件的支持度的历史变化规律的呈现和分析,得到可保留的规则。并帮助用户充分理解规则,合理运用规则。这项研究对知识发现的后处理与可实现性、实用性起着重要的作用。

关键词:知识发现;信息扩张机制;意外规则;支持度;可信度;充分性因子
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)12-0074-04

Research on Exceptional Rules Selecting and Comprehensibility Based on Information Increasing Mechanism

HU Jian[1], LUO Jia-guo[1], YANG Bing-ru[2]
(1. Jiangxi University of Science and Technology, Ganzhou 341000, China;
2. Beijing University of Science and Technology, Beijing 100083, China)

Abstract: With information increasing mechanism in the domain of knowledge discovery, a new processing method is proposed, that is the changes in trends analysition of three-matrics(rule's support degree; credibility and sufficiency factors). By analysing the history variation of the support degree of rule's antecedents and consequents, the rules are retained. Then help user to understand and reasonable use ruls. The research will play an important roles in post-processing, realizability and practicality of knowledge discovery.

Key words: knowledge discovery; information increasing mechanism; exception rule; support degree; credibility; sufficiency factors

0 引 言

知识发现内在机理中的信息扩张机制,其描述性定义为:当知识发现过程从一个抽象级向下一个抽象级,从一个固有数据库(知识库)向扩展数据库(知识库)过渡的时候,表现出来的的运行规律和处理方式。信息扩散原理是一种在样本不足的情况下,对样本应遵循的规律进行认识的模糊数据处理方法。文中提出的自动评价方法可在领域专家不介入的情况下,利用知识(规则)的可计算参数进行评价;并由信息扩散原理弥补参数相对不足的缺陷,得到规则参数的概率分布信息,据此客观地展现规则特征,从而实现规则评价。

文献[1~6]提出了各具特色的意外规则挖掘方法,它们的共同之处是利用了意外规则和普通规则是规则对的性质进行挖掘,但它们都没有对其进行深入的评价[7],而且通过研究发现同一条规则可以提供不同的信息,上述文献均没有对这个问题进行研究。事实上,因为规则是在事物发生、发展的一个时间点上得到的,受很多随机因素的影响,其提供的信息的价值就受到了质疑[8]。所以寻找一种方法分解各种可能、解释各种可能就成为一件非常有意义事情,其运用的分析手段可以平行地移植到其它规则的取舍和理解上。

文中提供一种方法——三度(规则支持度、规则可信度、规则充分性因子)变化趋势分析,再结合规则前件、后件的支持度的变化规律的呈现和分析,首先排除不正常数据,即噪音数据造成意外规则的情况;然后判断规则的发展进程中是否频繁满足文中提出的规则舍弃定理中的条件,由此得到可保留的规则。但同一条规则能提供给用户(决策者)的信息是不同的,通过三度变化规律的分析,得出对同一规则的不同理解,帮助决策者充分理解规则,合理运用规则。

这项研究对知识发现结果真正服务于用户起到了一定的作用。

收稿日期:2008-06-28
基金项目:国家自然科学基金资助项目(60675030)
作者简介:胡 健(1967-),男,副教授,博士,研究方向为数据挖掘与智能信息检索;杨炳儒,教授,博士生导师,研究领域为知识发现与智能系统,柔性建模与集成技术。

1　意外规则可能的变化趋势与三度分析相关性

伴随着规则的出现而存在的参数有规则前件支持度、后件支持度、规则支持度、规则可信度、规则充分性因子[9,10]。在考虑规则的历史表现时,上述的每一种参数都应考虑,否则就难免失之偏颇。但如果每一参数都考虑上升、下降、平行保持三种变化可能,那么上述 5 个参数的变化组合就为 $3^5=243$ 种。如果对其一一考察,无疑是笨拙的和难于完成的;但分析中又不应遗漏任一可能组合。所以下文首先给出规则支持度、可信度的可能变化;其次提出并证明 5 种参数之间应具有的关系,利用此关系排除 243 种组合中不可能出现的组合,余下 20 种组合。为了书写和阅读的方便,将具有共性的组合聚集在一个主题下,共给出 4 个主题;将 20 种组合分布在其中依次给出分析结论,并探索、解读各种可能情况,向用户提示规则中蕴涵的深层信息。

首先,分析支持度可能的变化趋势:意外规则的最小支持度和最小可信度为 eminsup 和 eminconf,显然,eminsup<minsup,而意外规则的支持度应在[eminsup, minsup]范围之内。所以 S_i 的变化有两种可能:从高至低,表现下降趋势;平行保持趋势。虽然,由低至高的可能存在,但所谓的高由于受到"支持度很小"和小于 minsup 的条件限制,低也受到大于等于 eminsup 的限制,所以变化趋势并不明显,可与 2 等同对待。可信度的趋势分析同理,存在两种可能:从低到高,表现上升趋势;平行保持趋势。

其次,给出"规则舍弃"的定理:假设 A 为规则前件,B 为规则后件,$r:A\rightarrow B$ 表示规则,S_A 表示前件支持度,Sr 表示规则支持度,其余类同。

命题 1:当满足如下条件之一时规则舍弃:

① $LS<1$ 即 $S_B>Cr$

② $LS_{A\rightarrow B}<LS_{B\rightarrow A}$ 即 $S_A>S_B$

说明:① 可由式(2) 证得;② 由文献[8] 提出,并已给出证明。

再次,阐明如下事实,揭示 5 个参数之间的关系。

命题 2:规则可信度 $C_r\geqslant$ 规则支持度 S_r(显见)。

定理 1:C_r,S_r 与 S_B 有一定关系,但不足以影响 S_B,进而影响 LS 的变化趋势,即当 C_r,S_r 表现出上文中的变化趋势时,LS 的变化是不一定的,可以上升、下降和平行不变。

定理 2:当 C_r 平行保持时,规则充分性因子 LS 与规则后件支持度 S_B 成反比。

定理 3:$0<S_r<S_A<C_r<1<LS, S_r<S_B$

证明:在数据库中 $P(B/A)$ 即 C_r,S_B 即 $P(B)$,S_r 即 $P(A\cup B)$

因 $LS_{A\rightarrow B}=\dfrac{P(B/A)\times(1-P(B))}{P(B)\times(1-P(B/A))}$

故如果 $P(B/A)$ 保持不变,则 $LS_{A\rightarrow B}$ 与 $P(B)$ 成反比,则定理 2 得证;如果 $P(B/A)$ 保持上升趋势,则因 $P(B)$ 变化不定,则 $LS_{A\rightarrow B}$ 变化不定,则定理 1 得证。

因 $C_r=\dfrac{S_r}{S_A}<1$　故 $S_r<S_A$ 同理 $S_r<S_B$,

因 $\mathrm{corr}=\dfrac{S_r}{S_AS_B}>1$　故 $S_A<C_r$ 则定理 3 得证。

2　意外规则的取舍和可理解性分析

基于以上事实,给出意外规则各参数历史变化的图形化表示形式的完备性讨论。

2.1　支持度保持平行,可信度保持平行

支持度保持平行,可信度保持平行见图 1。

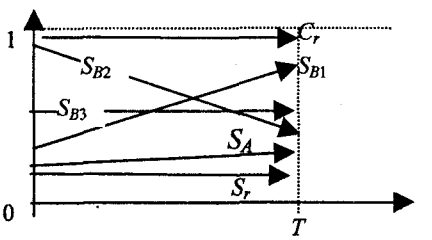

图 1　支持度保持平行,可信度保持平行

分析图 1(以商场销售记录为背景,其余类同):

(1) 支持度 S_r 保持平行,可信度 C_r 保持平行,说明规则的意外性稳定出现,排除了由于误操作等因素造成的意外情况,那么只需考察规则的 LS 的变化即可决定规则的取舍。由定理 2 知,LS 的变化取决于 S_B 的变化。

(2) 支持度 S_r 保持平行,可信度 C_r 保持平行,根据二者定义 S_A 保持平行。

(3) 在时刻 T,$LS>1$,假设临时知识库的规则已经经过'对规则取舍算法'Prara1(Prara1 是文献[8] 中提出的算法,根据文中证明的'对规则定理' 提出的对形如 $A\rightarrow B$ 和 $B\rightarrow A$ 这样的'对规则' 进行取舍的算法) 的操作,即 $S_A<S_B$。

(4) 在 T 时刻前,当 $S_B<S_A$ 时,或 $S_B>C_r$ 时,$LS<1$,规则舍弃。但如图 1 所示,当满足 S_r 平行,C_r 平行的条件时,上述情况的发生应仅为点状分布,因本节试图通过三度的变化过程分析规则,故个别点的情况可以忽略,则 S_B 的变化为图示的 S_{B1}、S_{B2}、S_{B3} 三种。

(5) 根据上述分析,首先规则的意外性表现稳定,其次 $LS>1$ 并始终大于其对规则,则规则应保留。

(6) 下面分析其可理解性:根据定理 2,S_{B2} 出现时 LS 上升,出现支持规则的特性,即 S_B 销售量的下降,使它的出现更加依赖 A 的出现,可充分信任规则,但需做出减少 B 的进货量的类似决策;如 S_{B1} 出现,则用户需注意,虽然事务 A 决定了事务 B,但 B 的畅销使其对 A 的依赖越来越小。S_{B3} 的出现说明各参数表现平稳,T 时刻计算出的各参数可信。

2.2 支持度下降,可信度保持平行

支持度下降,可信度保持平行见图 2。

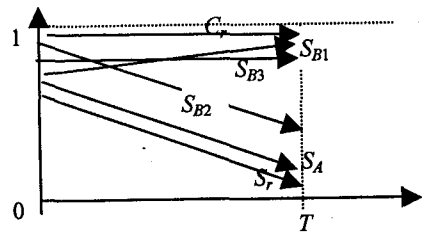

图 2 支持度下降,可信度保持平行

分析图 2:

(1) 从图 2 中可以看出规则在 T 时刻表现出来的意外性是在数据渐增过程中发展而来的,原来此规则的各项参数表现良好,是一条强规则。其平稳过渡得来的意外性说明非噪声数据引起,故可进行下阶段分析。

(2) 与图 1 的分析相同,因为可信度 C_r 在一个较高的水平保持平行,出现定理 1 ① 的情况仅可能是点状分布,就变化过程而言,$LS > 1$ 会在数据库渐增过程中保持。

(3) 支持度 S_r 下降,可信度 C_r 保持平行,根据二者定义,S_A 下降,如图。

(4) 根据定理 3,S_B 的变化如图所示。其中 S_{B1}、S_{B3} 出现时,即使小于 S_A 的情况发生,也不会演变成一种趋势,则规则保留。

(5) 当 S_{B2} 出现时,如果 $S_{B2} > S_A$ 较多发生时,应保留规则,并且相对于相反的情况其更常见。相反,$S_{B2} < S_A$ 较多发生时,此情况并不常见,但其出现则意味着需舍弃规则。

(6) 下面分析其可理解性:规则支持度的下降说明,事务 A 与事务 B 的共同出现的情况越来越少发生,但高可信度和渐升的 LS(S_{B2} 出现时,在保留规则的前提下) 仍可保证 A 与 B 是有很大的关联并且 B 越来越依赖 A,所以可充分信任规则,但应考虑作为物品的 A、B 已稳定地出现销量下降的情况。如 S_{B1} 出现,则用户需注意 B 对 A 的依赖越来越小,并且 A 的销量在下降。S_{B3} 出现说明 B 对 A 的依赖保持不变,是 A 的销量的下降使得 A、B 共同出现的次数减小,而 B 的销量仍很稳定。

2.3 支持度保持平行,可信度上升

支持度保持平行,可信度上升见图 3。

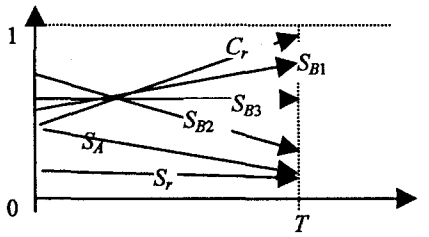

图 3 支持度保持平行,可信度上升

分析图 3:

(1) 支持度 S_r 保持平行,可信度 C_r 上升,根据二者定义,S_A 下降,如图。

(2) 如图所示,如果 $S_A > C_r$ 的情况频繁发生,即 $LS < 1$ 频繁出现,则规则舍弃;否则 $LS > 1$。

(3) 如果 S_B 频繁地小于 S_A 则规则舍弃。

(4) 排除了(2)、(3) 两种情况,则规则可以保留。

(5) 下面分析其可理解性:在规则保留的情况下,也要针对不同情况运用这条规则。即 S_{B2} 出现时,结合 C_r 上升,LS 必上升,则 A 与 B 的关系越来越紧密,并且 B 越来越依赖 A,但需注意规则的意外性是由 S_A、S_B 分别持续下降造成的;S_{B3} 出现时,结合 C_r 上升,LS 必上升,则情况与上同;S_{B1} 出现时,结合 C_r 上升,LS 变化不定,此时则需通过分库计算 LS,观察其变化,再加深对规则的理解。

2.4 支持度下降,可信度上升

支持度下降,可信度上升见图 4。

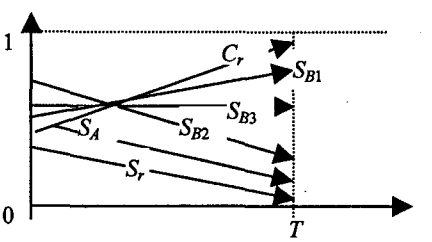

图 4 支持度下降,可信度上升

分析图 4:

支持度下降,可信度上升时,S_A 的降幅是远大于 S_r 的,但图 4 中 S_r 和 C_r 的格局没有为 S_A 留出足够的下降空间,因此图 4 出现的可能性不大。如出现,则图 4 与图 3 的取舍条件是完全一样的,在理解上也大体相同,但确需注意到 A 与 B 共同购买量在减少。

除图 1 ~ 4 所示情况,各参数还可表现出无规律分布的情况,此时在 T 时刻挖掘出的意外规则是由不明原因的随机因素引起的,无法预知在下一时间段内规则表现如何,则规则更不可轻信,笔者建议舍弃。

从以上分析可看出,只有图 1 所示规则才是人们

心目中真正期待的意外规则,而其它各种情况,其意外性只在 T 时刻和其有限的时间邻域内保持,但终将顺着原发展趋势发展下去,而失去了那些意外规则所具有的特征。所以只有图 1 所示规则才能长期遵循,其余情况只能在一个时间段内相信。如果图 1 中出现参数大幅的跳动,排除噪声干扰的情况后,则认为此时有可能发生了规则突变,为此正在进行知识发现过程中的突变理论的研究。

3　实验验证

当挖掘出意外规则时,必须对数据库 DB 进行分库[11]。分库的方法有三种:第一种:假定将时间 T 分割成 n 段 T_i, trac_ time i 表示事件发生时间($i = 1,2,\cdots,n$), T_i = [trac_ time 1, trac_ time i],并且对任意 $i \neq j, i, j = 1,2,\cdots,n$,如果 $i < j$,则 trac_ time i < trac_ time j,并且 T = [trac_ time 1, trac_ time$n$];第二种:根据 T_i 将数据库逻辑地分成 DB_i 个,每个 DB_i 与 T_i 对应。在这里并不强调一定采用第二种分库方案,针对没有时间属性的某些非商业数据库(往往是多值型的),提出第三种分库方案;第三种:将 DB 逻辑地分成 n 个 DB_i 满足条件,当 $i < j$ 时, $DB_i \subset DB_j$,并且 $DB_n = DB$。

经过分库,得到 n 个子库,在 n 个子库上分别计算欲分析之意外规则的支持度 S_i,可信度 C_i 和充分性因子 $LS_i, i = 1,2,\cdots,n$。按照第三种的分库方案将数据库分成 10 个子库,并编制程序计算上述各参数,并用二维曲线图的形式(见图 5)表示出来。

如图 5 所示,正是图 1 中 S_{B2} 出现时的情况,则根据图示,结合上文分析,用户即可决定保留此条规则。

4　结束语

文中属知识发现内在机理研究中信息扩散机制的一个专题。针对意外规则难取舍问题提出了解决方案:给出了舍弃规则的条件;阐明了三度及规则前件、后件之间各种繁杂的关系,及其相互制约的条件;利用前两者的结果,从规则各参数变化过程的角度分析规则得出结论;并编程实现了这一分析方法,得到了与理论分析相符的应用结果。

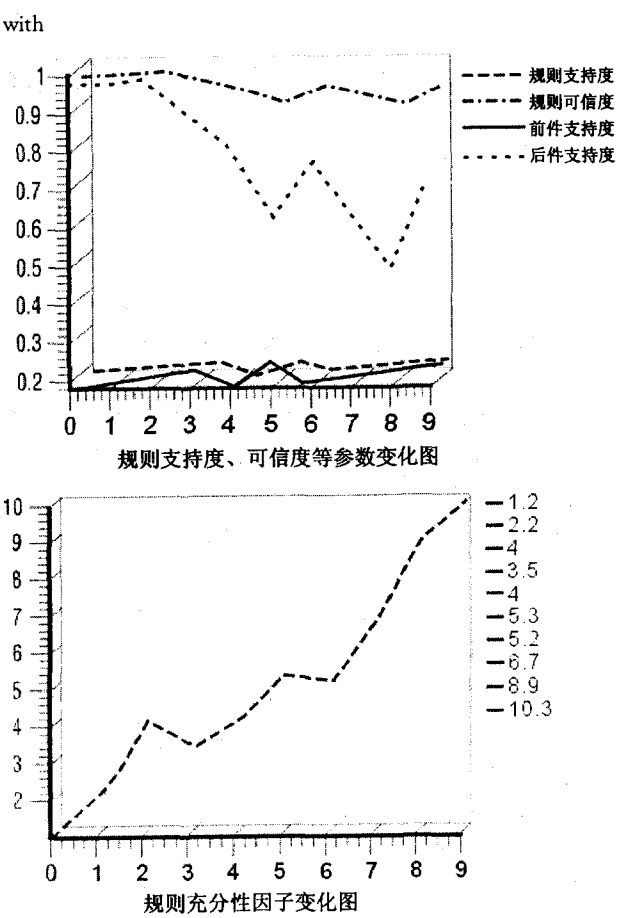

图 5　基于信息扩张机制的意外规则分析系统

参考文献:

[1] Liu Huan, Lu Hongjun, Feng Ling, et al. Efficient Search of Reliable Exception[C]//In Proc. of PAKDD－99, Third Pacific－Asia Conference. Beijing, China: [s. n.], 1999: 194－203.

[2] Hussain F, Liu H, Suzuki E, et al. Exception rule mining with a relative interestingness measure[C]//In Proc. Fourth Conf. PAKDD－00. London: Springer Verlag, 2000: 86－97.

[3] Suzuji E, Shumura M. Exceptional Knowledge Discovery in Databases bases on Information Theory[C]//In Proc. of KDD－96. AAAI. Portland, Oregon, USA: [s. n.], 1996: 275－278.

[4] Suzuki E. Autonomous Discovery of Reliable Exception Rules [C]//In Proc. of KDD－97. Newport Beach, CA, USA: [s. n.], 1997: 259－262.

[5] Suzuki E, Tsunnoto S. Evaluating Hypothesis－driven exception－rule with medical data sets[C]//In proc. Fourth Conf. PAKDD－00. Kyoto, Japan: [s. n.], 2000: 208－211.

[6] 孙海洪,夏克俭,杨炳儒,等.一种挖掘意外规则的快速算法[J].计算机工程与应用,2001,37(19):49－52.

[7] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范　明,孟小峰等译. 北京:机械工业出版社,2001.

[8] 周　颖. 对规则取舍问题的研究[J]. 计算机科学,2003,30(5):102－104.

[9] 欧阳为民,蔡庆生. 在数据库中发现具有时态约束的关联规则[J]. 软件学报,1999,10(5):527－532.

[10] 欧阳为民,蔡庆生. 数据库中的时态数据发掘研究[J]. 计算机科学,1998,25(4):60－63.

[11] 施平安,陈文伟,黄金才. 关联规则时间适用性及其发现方法[J]. 计算机应用研究,2001,18(6):18－20.