对规则取舍问题的研究*)

周 颖
(北京科技大学信息工程学院 北京100083)

A Study of a Pair of Rules' Acceptant or Rejective Problem

ZHOU Ying
(Information and Engineering Institute, the University of Science and Technology, Beijing 100083)

Abstract This paper establishes a theory architecture to accept or reject or reserve synchronously a pair of rules look like A→B and B→A by analyzing correlation and sufficiency gene, and presents its algorithms. Prara1 reduces domain specialist workload, realiazes evaluative goal of decreasing people's limitation. Prara2 also reduces workload of mining-evaluating algorithm.

Keywords Sufficiency gene, A pair of rules, Auto-evaluate

当前KDD发展的主流是寻求在各类数据库和应用问题的背景下高性能、高扩展性的发掘算法[1]。但知识(规则)评价越来越多地受到更多学者的关注和重视,然而涉及评价的文献还是少之又少。现实世界是千变万化的,挖掘出的规则只能在一定的条件下才能成立,评价的研究正逐步朝着解脱领域专家,并摆脱人的局限的方向进行。对关联规则,人们相继提出了继支持度、可信度之后的相关性度量[2]或充分性因子[3]的评价方法。通过它们可以找到强关联规则中的有趣规则,删除对用户会产生误导的规则。但即使经过了如此的评价,知识库中仍然会存在这样的规则,如'珍珠→盒子'和'盒子→珍珠',显然后者是不合理的,目前的评价方法只能通过领域专家的人工操作进行删除,还没有就此问题形成可靠的理论依据和实用、简便的可实现算法,本文正是就此问题进行研究,提出解决方案。

假设用A→B表示规则(A,B均为知识合节点[1],下同),记为r_1;规则B→A,记为r_2,则此时应分两种情况考虑:1.A与B的销售确实互相依存,共消共长;2.珍珠带动盒子的销售,相反的情况并不存在。本文即试图通过对相关性度量、充分性因子的分析建立对形如A→B和B→A这样的对规则进行取舍或同时保留的理论体系,并构造其实现算法。下文中称规则A→B和规则B→A为对规则。

1. 相关性度量与充分性因子的关系:

1.1 充分性因子LS[3]

主观Bayes方法中,每条规则的表示形式[4]是

IF A THEN (LS,LN) B (P(B))

其中,P(B)是B的先验概率,在关联规则中就是B的支持度;LS∈[0,+∞]称为充分性因子,它反映了证据A为真对结论B的影响程度。LS表示如下:

$$LS=\frac{P(B/A)\times(1-P(B))}{P(B)\times(1-P(B/A))} \quad (1)$$

其中,P(B/A)是条件概率,P(B)是B的先验概率,在关联规则挖掘过程中,可以从数据库中得到P(B),P(B/A)等的概率值。从式(1)中可以看出LS的意义:

(1) 当LS=1时,P(B/A)=P(B),这表明A与B无关;

(2) 当LS>1时,P(B/A)>P(B),这表明由于A所对应的证据存在,增大了B为真的可能性,而且LS越大,P(B/A)就越大,即A对B为真的支持越强。当LS→∞时,P(B/A)→1,表明由于A的存在,将导致B为真;

(3) 当LS<1时,P(B/A)<P(B),这表明由于证据A的存在,将导致B为真的可能性下降;

(4) 当LS=0,P(B/A)=0,这表明由于证据A的存在,将使B为假。

由上述讨论可以看出,只有在(2)的情况下,LS>1,即P(B/A)>P(B)时,由于A所对应的证据存在,增大了B为真的可能性,有用的关联规则的LS都应该大于1,也即只有关联规则的可信度P(B/A)大于先验概率P(B),才说明A的出现对B的出现有促进作用,也说明了它们之间有某种程度的相关性。反之,如果充分性因子LS不大于1,则此关联规则也就没有意义了,可以删除。

充分性因子描述规则的条件对结论的影响力的大小,充分性因子越大,说明结论受条件的影响越大(性质1)。

1.2 相关性度量

如果项集A的出现独立于B的出现,则P(A∪B)=P(A)P(B);否则,项集A和B作为事件是依赖的和相关的。这个定义容易推广到多于两个项集。A和B的出现之间的相关性通过计算

$$corr_{A,B}=\frac{P(A\cup B)}{P(A)P(B)} \quad (2)$$

来度量。如果(2)式的值小于1,则A的出现和B的出现是负相关的。如果结果值大于1,则A和B是正相关的,意味着每一个的出现都蕴涵另一个的出现。如果值等于1,则A和B是独立的,它们之间没有相关性。

1.3 二者的关系

由于式(2)等价于$P(B|A)/P(B)$,则显然相关性度量也具有1.1节中描述的充分性因子的性质,即1.1节中列举的(1)~(4)条。但相关性度量只能判断A、B的相关性,却不能象充分性因子那样说明条件对结论的出现有促进作用并用数学的方法表示这种作用。所以,充分性因子蕴涵了相关性度量。

*)国家自然科学基金重点项目(69835001),北京市自然科学基金(4022008)。周 颖 博士生,主要研究方向为知识发现。

2. 对规则定理

因为 $corr_{A,B}=corr_{B,A}$，也就是说如果 A 与 B 是正相关的，则 B 与 A 也是正相关的，即当 $corr_{A,B}>1$ 时 $corr_{B,A}>1$ 同时成立，所以，由1.3节的分析知 $LS_{A\to B}>1$ 与 $LS_{B\to A}>1$ 也能同时成立（正因于此，对规则才会被同时保留下来）。基于此给出对规则定理：

定理1（对规则定理） 假设有一对规则 A→B 和 B→A，项集 A 的支持度为 $P(A)$，项集 B 的支持度为 $P(B)$，那么 $P(A)\leqslant P(B)$ 的充分必要条件是 $LS_{A\to B}\geqslant LS_{B\to A}$。

证明：先证充分性

$\because P(B/A)=P(A\cup B)/P(A)$

$$\therefore LS_{A\to B}=\frac{P(B/A)\times(1-P(B))}{P(B)\times(1-P(B/A))}=\frac{P(A\cup B)\times(1-P(B))}{P(B)\times(P(A)-P(A\cup B))}$$

$$LS_{B\to A}=\frac{P(A/B)\times(1-P(A))}{P(A)\times(1-P(A/B))}=\frac{P(A\cup B)\times(1-P(A))}{P(A)\times(P(B)-P(A\cup B))}$$

$\because LS_{A\to B}\geqslant LS_{B\to A}$

$$\therefore \frac{P(A\cup B)\times(1-P(B))}{P(B)\times(P(A)-P(A\cup B))}\geqslant\frac{P(A\cup B)\times(1-P(A))}{P(A)\times(P(B)-P(A\cup B))}$$

$\because P(A)>P(A\cup B)$

$\therefore P(A)\times(P(B)-P(A\cup B))\times P(A\cup B)\times(1-P(B))\geqslant P(A\cup B)\times(1-P(A))\times P(B)\times(P(A)-P(A\cup B))$

经化简可得：$P(A)\leqslant P(B)$

再证必要性

以上证明过程沿原路返回，即得。证毕。

3. 对规则取舍算法 Prara1(a Pair of Rules' Acceptant or Rej-ective Algorithm 1)

此算法利用了1.1节中的性质1和第2节中证明的对规则定理，在 A，B 中选择支持度较小者为前件组成规则，例如：如 $P(A)<P(B)$，则规则 A→B 保留，规则 B→A 被删除。由对规则定理可知：当 $P(A)=P(B)$ 时，$LS_{A\to B}=LS_{B\to A}$。这说明 A 与 B 共消共长，互相依存，不分彼此，则对规则 A→B 和 B→A 均有保留价值，均应保留。这是理想状态，还应考虑现实世界并非非此即彼：数据录入的错误、事件发生的随机性等等，都可能使原本共消共长的 A 与 B 发生偏差，表现为 P(A)，P(B) 存在差值；另外也不能不考虑用户担心丢失有趣规则的心理，所以在算法的具体实现时由用户给定差值阈值，记为 a。当 $|P(A)-P(B)|\leqslant a$ 时，则保留对规则 A→B 和 B→A；反之按性质1，保留充分性因子大者，即根据对规则定理，保留前件支持度小者。下面是 Praral 算法描述：

输入：经相关性度量分析后的规则集

输出：删除对规则中无趣规则后的规则集

Step1：for i:=1 to n do //n 是输入的规则集中的规则数

Step2：If 第 i 条规则在输入规则集中存在对规则 then

Step3：if abs(Supp(第 i 条规则前件)－Supp(第 i 条规则后件))>a then //a 是差值阈值

Step4： if Supp(第 i 条规则前件) > Supp(第 i 条规则后件) then //Supp(A)表示项集 A 的支持，即前文中的 P(A)，Supp(A)已经在挖掘算法中得到

Step5： delete 规则集中当前记录；

Step6：结束

有些挖掘算法在挖掘过程中可能已经通过各种剪枝方法，将对规则中的一个规则剪掉了，那么 Prara1算法考虑到了这种情况，没有将对规则仅存的这条规则剪掉。如果挖掘算法中保留了所有对规则，则 Prara1算法需要去掉 Step2。

关于差值阈值 a 的问题，可以由领域专家在系统维护的参数维护中输入，程序运行过程中不能更改；也可以在 Prara1 算法运行前由用户输入，真正删除记录前可以修改，不同语言提供了几种不同的处理方法，如对记录虚删除，整个规则集运行完毕后，如不需修改 a 值，再真实删除，否则重新运行 Prara1算法；或建立规则集的视图，在其上删除，确认后再在规则集上真实删除等等。

4. 实例运行

在某地区高考数据库上，有属性：学号，学生姓名，性别，录取校名，高考总成绩……，其中高考总成绩是连续属性，需要进行离散化，程序中采用的离散化方法是在笔者的另一篇论文《基于语言场理论的连续属性离散化方法及实现》（已投计算机科学）中介绍的 DCL 算法，高考总成绩被定义为5个语言值：很低，低，一般，高，很高。录取校名采用"一一对应"方法，录取校名一一对应为语言值，离散化后的部分记录如表1所示（只显示与说明问题有关的属性）。

表1 挖掘数据库中部分属性上的部分记录

记录编号	录取校名	高考总成绩
105	××科技大学	高
106	××理工大学	高
107	××师范大学	高
108	××语言大学	很高
109	××科技大学	低
110	××科技大学	高
111	××科技大学	高
112	××理工大学	高
113	××科技大学	很高

运行挖掘算法后，得到'××科技大学→高考总成绩高'（记为 r_1）和'高考总成绩高→××科技大学'（记为 r_2）这样的对规则。运行 Prara1算法，规则 r_1 保留，r_2 被删除。

可以想象，数据库中××科技大学支持度一定小于高考总成绩高的支持度，因为数据库中还有很多学校录取的学生高考总成绩也是高。规则 r_2 没有保留的价值，也是显而易见的。

5. 对规则取舍算法 Prara2

Prara1算法虽然基于对充分性因子的讨论，但具体实现上却完全不涉及充分性因子，只是利用了挖掘算法运行时已经得出的项集的支持度进行比较，即可删除挖掘结果中近一半的无趣规则。对规则取舍原则更大的用途还不仅于此，可以将它用在从大项集中生成规则之前，即先比较前、后件的支持度，构造以小支持度的项集为前件的规则，即对已挑选出来做前件和后件的项集的支持度进行比较，如下：

Step1：if abs(Supp(第 i 条规则前件)－Supp(第 i 条规则后件))>a then //a 是差值阈值，处理同 Prara1

Step2: if Supp(第i条规则前件) ≤ Supp(第i条规则后件) then

Step3:计算规则可信度

那么因为近一半的规则不用计算其可信度,更不用进行评价,所以不仅不会增加原挖掘一评价算法的负担,还在一定程度上减少了它的时空复杂性。此算法命名为 Prara2。

以小支持度的项集为前件的规则其可信度也比以它为后件的规则的可信度大,但是不能以哪条规则的可信度大为依据在对规则中决定取舍,因为可信度并不具备充分性因子的性质,即1.1节中的性质1,而且充分性因子是由可信度和规则后件支持度共同决定的。

结论 用相关性度量识别关联规则的相关性,已经被广泛地采用,但它不能识别作为规则前件和后件的两个项集中哪一个对另一个的出现具有更大的促进作用;可信度分析也只能说明对规则中哪一个的可信度高,仅此而已;孙海洪博士在他提出的 QAR_SQL 算法中采用了充分性因子进行评价,但也只是将相关性定量地给出了。本文利用相关性度量和充分性因子的关系和后者的性质提出了对规则取舍问题的解决方案,它的意义在于减少了领域专家的工作量,些微地推进了自动评价技术的发展,在一定程度上解决了 KDD 主流发展中存在的问题之一——领域专家的局限。这方面的工作还将继续进行。

参考文献

1 杨炳儒. 知识工程与知识发现. 冶金工业出版社,2000

2 Han Jiawei. Micheline Kamber. Data Mining: Concepts and Techniques. 高等教育出版社,2001

3 孙海洪. KDD 算法和启发型协调器的理论研究及其应用:[博士论文]. 北京科技大学,2001

4 王永庆. 人工智能原理与方法. 西安交通大学出版社,1998. 162~171

(上接第83页)

文件,但由于在数据收集中代理服务器和缓冲技术的影响,产生的会话文件无法做到很准确。

Episode 识别是一个有选择性的预处理步骤,一般是放在必处理步骤之后,是由 W3C 定义的一个用户会话的语义子集。

4.2 模式发现

一但用户会话和事务已被识别就可采用以下技术进行挖掘。

(1)*路径分析* 判断在一个 Web 站点中最频繁访问的路径,其它的相关路径可通过路径分析得出,利用这些信息还可以改进站点的结构设计。

(2)*关联规则和序列模式的发现* 使用关联规则发现相关性,利用相关性更好地组织站点内的 Web 空间,实行有效的市场战略。序列模式的发现:能够便于预测用户的访问模式,有助于开展这种模式的有针对性的服务。

(3)*分类和聚类* 分类规则可以识别一个特殊群体的公共属性的描述,并可以用来分类新的市场战略。聚类规则以概率分析为基础,发现客户访问网站的整体分布情况。

4.3 模式分析

模式分析的方法有:联机分析、可视化、知识查询和信息过滤。首先用模式分析工具将抽象的使用模式以直观、容易理解的方式展现给分析者,然后分析者利用知识查询语言,根据需要对挖掘过程加以限制,得到感兴趣的使用模式。比如限定某一领域进行挖掘,然后就这一领域挖掘出来的使用模式进行分析,得出感兴趣的结果。

信息过滤分两部分:objective 过滤和 subjective 过滤。objective 过滤处理用不同模式发现关联的数值型度量的变化,比如:支持度和兴趣度;subjective 过滤是用来处理使用挖掘通过分析网站内容和结构而形成访问网页的可信任度。对于 Web 使用挖掘,设想用网站结构和内容作为网站设计者的领域知识,在网页之间进行链接以提供这些页面的关联支持,那么在网页之间的拓扑链接越强,这些网页一起被访问的可信度也就越高。类似地,在同一个内容簇或同一类里的页面被认为在一起被访问的可信度远远大于不同簇或不同类中的页面。

结束语 本文进一步强调了 Web 的结构和内容不仅仅对使用挖掘有影响,而且是至关重要的和密不可分的。它是使用挖掘处理算法的重要数据源,贯穿整个使用模式发现的全过程,并且能为模式分析中的用户行为的预测提供信息。进一步讲,网站结构和内容的挖掘结果又可以作为很重要的数据源进行网页的分类和聚类,从而提高 Web 使用挖掘的效率。

参考文献

1 Cooley R ,Tan P-N, Srivastava J. Websift: The Web site information filter system. In WEBKDD,San Diego,CA,1999

2 Liu Lizhen. The Research of Web Mining. In: The 4th World Congress on Intelligent Control and Automation,2002

3 Cooley R,Mobasher B,Srivastava J. Data preparation for mining world wide Web browseing patterns. Knowledge and Information Systems,1999

4 Linoff G S,Berry M J A. Mining the Web,2001

5 Mena J. Data Mining Your Website,1999

6 Fayyad U,et al. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11)

7 Hahn U, Schnattinger K. Deep knowledge discovery from natural language texts. In: Proc of the 3rd Int'l Conf. Knowledge Discovery and Data Mining ; Newport Beach, 1997

8 Wang Wei Qiang. Text Ming on the Internet Computer Science. 2000

9 Wang Ji Chen. Research on Web Text Mining. Journal of computer Research&Development, 2000,37(5)

10 Chen En Hong. Web Usage Mining: Discovering User Behavior Patterns From Web Data. Computer Science, 2001,28(5)

11 Yang Xiao Hua. Hyperlink Structure Mining of Web Sites. Computer Project&Application, 2001,8

12 Wang Shi. Web Mining. Computer Science, 2000,27(4)

13 Wang xiao ya. ,Web Usage Mining:[PH. D thesis]. 2000,3

14 Liu Jun. Web Usage Mining:[Master thesis]. 2000