文章编号:1005－0523(2004)05－0023－03

基于包含度的事务数据库关联规则挖掘

罗来鹏,刘二根,胡新根,王广超

(华东交通大学 基础科学学院,江西 南昌,330013)

摘要:通过包含度刻画了关联规则信任度概念及讨论了它与Rough set精度的关系,特别地通过包含度的性质推导了信任度的增量计算.

关 键 词:包含度;信任度;精度

中图分类号:U279.3+24 **文献标识码**:B

0 前 言

在事务数据库关联规则的挖掘中存在大量不确定性问题,如何将不确定性推理中有关理论应用到事务数据库关联规则挖掘一直以来是一个有待解决的问题.

张文修教授[2]等根据各种不确定性推理提出了包含度理论,它作为不确定性研究的方法已被广泛应用.它是一种描述不确定性关系的有效度量方法,在人工智能、专家系统和模糊集理论等领域有着重要的应用.本文通过包含度刻画了关联规则信任度概念,及它与Rough set精度联系,并由该种包含度所具有的性质推导了事务数据库关联规则信任度的增量计算.

1 包含度

1.1 包含度的定义

包含度[2]是指一个集合 A 包含于一个集合 B 的程度 $D(B/A)$.它是一种处理不确定性研究的方法,是对"包含关系"的扩充,从而包容了"关系"的不确定性,满足以下公理:

公理1 $0\leqslant D(B/A)\leqslant 1$

公理2 $A\subset B$ 时,$D(B/A)=1$

公理3 $A\subset B\subset C$ 时,$D(A/C)\leqslant D(A/B)$

公理4 $A\subset B$ 时,对于任意 C 有 $D(A/C)\leqslant D(B/C)$

上面四条公理是针对经典子集定义的,公理1是包含度的规范化,包含度在[0,1]中取值;公理2是包含度和经典包含的协调性,经典包含关系是包含度为1的特殊情况;公理3、公理4是包含度的单调性.粗略地说,一个较小集合比较容易包含于一个比较大的集合里边.

若 A、B 是经典有限集,通常情况下,用以下方法来定义一个集合 B 包含集合 A 的程度:

$$D(B/A)=\begin{cases}|A\cap B|/|A| & A\neq\phi\\ 1 & A\neq\phi\end{cases}$$

这里 $|\cdot|$ 表示集合的基数,它在数据挖掘方面具有重要的应用价值.

下面都按此定义来讨论它的有关性质与应用.

1.2 一种重要包含度的性质

对于有限集 A、X、X_1、X_2,上述定义的包含度具有如下的一些性质:

性质1 $D(A/A)=1$,$D(\phi/A)=0$,一般有 $0\leqslant D(X/A)\leqslant 1$.

收稿日期:2004－05－20

作者简介:罗来鹏(1973－),男,江西吉水人,助教,研究方向:智能信息处理、数据挖掘.

证明略.

性质 2 $D(X/A)=1-(1/|A|)(|A\cup X|-|X|)$.

证明 因为$|A\cap X|=|A|+|X|-|A\cup X|$,所以$|A\cap X|/|A|=1+(1/|A|)(|X|-|A\cup X|)=1-(1/||A|)(|A\cup X|-|X|)$.

同理可证明下面的性质

性质 3 $D(X/(X_1\cap X_2))=1-(1/|X_1\cap X_2|)(|(X_1\cap X_2)\cup X|-|X|)$.

性质 4 $D(X/(X_1\cap X_2))=1-(1/|X_1\cup X_2|)(|X_1\cup X_2\cup X|-|X|)\leqslant(1/||X_1\cup X_2|)(|X_1||D(X/X_1)+|X_2|D(X/X_2))\leqslant D(X/X_1)+D(X/X_2)$.

性质 5 如果集合 $X_1\subseteq X_2\subseteq U$ 和任意集合 X,则 $D(X_1/X)\leqslant D(X_2/X)$.

性质 6 如果集合 $X_1,X_2\subseteq U$,则对于任意集合 X,有 $D((X_1\cap X_2)/X)=D(X_1/X)+D(X_2/X)-D((X_1\cup X_2)/X)$, $D(X_1/X)+D(X_2/X)\geqslant D((X_1\cup X_2)/X)\geqslant 0$.

性质 7 $D((X_1\cap X_2)/X)=|X_1\cap X_2\cap X|/|X|=(|(X_1\cap X_2)\cap X|/|X_1\cap X_2|)/|X|)=D(X/(X_1\cap X_2))(|X_1\cap X_2|/|X|)$

性质 8 $D((X_1\cup X_2)/X)=D(X_1/X)+D(X_2/X)-D((X_1\cap X_2)/X)=D(X_1/X)+D(X_2/X)-D(X/(X_1\cap X_2))(|X_1\cap X_2|/|X|)$. 显然,如果 $X_1\cap X_2=\phi$,则 $D((X_1\cup X_2)/X)=D(X_1/X)+D(X_2/X)$.

2 包含度与信任度、Rough Set 精度

在文[1]中关联规则 $X\rightarrow Y$ 的信任度定义为:$\text{confidence}(X\rightarrow Y)=P(Y|X)=|X\cap Y|/|X|=D(Y/X)$. 它体现的是包含 X 事务数集中包含 Y 事务数集的程度. 这里对于规则 $X\rightarrow Y$ 本身而言,X,Y 是项目集,但是对于式 $|X\cap Y|/|X|$,X,Y 是包含 X、Y 项目集的事务集,显然它是一种包含度.

在 Rough Set 中,对于有限论域 U,等价关系 R 和非空集合 X,则关于 X 的精度[3]为:

$$\mu_R(X)=|\cup\{Y\in U/R:Y\subseteq X\}|/|\cup\{Y\in U/R:Y\cap X\neq\Phi\}|$$

其中,U/R 是 R 的等价类簇.

设 $A=\cup\{Y\in U/R:Y\cap X\neq\Phi\}$, $B=\cup\{Y\in U/R:Y\subseteq X\}$,因为 $Y\in U/R$, $X\subseteq U$, $A=\cup\{Y\in U/R:Y\cap X\neq\Phi\}$,所以 $X\subseteq A$. 又 $B\subseteq X$,有 $B\subseteq A$,所以,$B=A\cap B$. 于是:

$$\mu_R(X)=|\cup\{Y\in U/R:Y\subseteq X\}|/|\cup\{Y\in U/R:Y\cap X\neq\Phi\}|=|B|/|A|=|B\cap A|/|A|$$

它也是一种包含度,而且它的计算形式与事务数据库中所定义的信任度是一样的,这说明从包含度这一数学本质上上述两种定义是相同的. 事实上规则 $X\rightarrow Y$ 的信任度等价于由 X 计算 Y 的精确度. 因此,在一些问题上可以将挖掘关联规则方法和分类规则方法融合在一起进行考虑.

3 信任度的增量计算

3.1 信任度的增量计算

信任度增量定理:设 U 是对象的有限集合,$D(X/A)=|A\cap X|/|A|$ 是集合 X 与集合 A 间的包含度,$A\subseteq U$, $X\subseteq U$. 数据库增加新数据记录以后,集合 $A'=A\cup\triangle A$,其中 $\triangle A$ 是 A 的增量部分,集合 $X'=X\cup\Delta X$,其中 ΔX 是 X 的增量部分,则 $D((X\cup\Delta X)/(A\cup\Delta A))=\eta D(X/A)+(1-\eta)D(\Delta X/\triangle A)$. 这里, $\eta=|A|/|A\cup\triangle A|$.

证明 由性质 4、8 有:

$$D((X\cup\Delta X)/(A\cup\triangle A))=2-(1/|A\cup\triangle A|)[|A\cup\triangle A\cup X|-|X|+|A\cup\triangle A\cup\Delta X|-|\triangle A||]=[|(A\cup\triangle A)\cap X|+|(A\cup\triangle A)\cap\Delta X|]/|A\cup\triangle A|.$$

事实上,根据性质 8 有:

$$D((X\cup\Delta X)/(A\cup\triangle A))=D(X/(A\cup\triangle A))+D(\Delta X/(A\cup\triangle A))-D((X\cap\Delta X)/(A\cup\triangle A))=D(X/(A\cup\triangle A))+D(\Delta X/(A\cup\triangle A))$$

因为$|X\cap\Delta X|=|\phi|=0$,所以 $D((X\cap\Delta X)/A)=0$, $D((X\cup\Delta X)/(A\cup\triangle A))=0$,

同样$|\triangle A\cap X|=0$, $|\triangle X\cap A|=0$, $|A\cap\triangle A|=0$.

若令 $\eta=|A|/|A\cup\triangle A|(0\leqslant\eta\leqslant1)$,由 $A\cap\triangle A=\phi$,则$|A\cup\triangle A|=|A|+|\triangle A|$,有$|\triangle A|/|A\cup\triangle A|=1-\eta$,得:

$$D((X\cup\Delta X)/(A\cup\triangle A))=\eta D(X/A)+(1-\eta)D(\Delta X/\triangle A).$$

3.2 进一步的结论

若令 $c=\text{confindence}(A\rightarrow X)$表示数据增量前规则的信任度,$\triangle c=\text{confidence}(\triangle A\rightarrow\Delta X)$为增量部分规则的信任度,$c=\text{confidence}(A'\rightarrow X')$为增量后规则的信任度,其实它们表示相同一条规则 $A\rightarrow X$,并设 c 为用户给定的最小信任度,则有以下推论:

推论 1:若 $c=1$, $\triangle c=1$ 则 $c'=1$.

推论 2:若 $c\geqslant \underline{c}$, $\triangle c\geqslant \underline{c}$,则 $c'\geqslant \underline{c}$.

推论 3:若 $c<\underline{c}$,$\triangle c<\underline{c}$,则 $c'<\underline{c}$.

由推论 2 可知,若一有效规则(这里仅从信任度而言),它的增量部分使得规则仍为有效,则该规则仍为有效的.

4 小　结

本文通过包含度分析了关联规则中的信任度及它与 Rough set 中精度的联系,发现了他们的基本计算在数学上是统一的,这为研究分类规则和关联规则的统一挖掘方法提供了一定的思路,尤其重要的是由于包含度已成为不确定性推理中一种重要的方法,这将为不确定性中的相关理论应用到事务数据库关联规则挖掘提供了可能.这种统一对于综合数据挖掘方法、开发数据挖掘语言具有重要的意义.另外,所推导的信任度的增量计算对于算法设计与分析也有一定参考价值.

参考文献:

[1] Jiawei Han, Micheline Kamber 著, Data Mining Concepts and Techniques[M]. 范明,孟小峰,等译. 北京:机械工业出版社(第一版),2001.

[2] 张文修,梁怡. 不确定性推理原理[M]. 西安:西安交通大学出版社,1996.

[3] Pawlak Z. Rough sets—Theoretical Aspects of Reasoning about Data[M]. Dordrecht:Kluwer Academic Publishers,1991.

[4] 梁吉业,等. 包含度与粗糙集数据分析中的度量[J]. 计算机学报,2001, 5(24).

[5] 张梅,李怀祖. 基于包含度方法的知识发现[J]. 系统工程理论方法应用,2001,4(10).

Mining of Association Rules of Transaction Database Based on Including Degree

LUO Lai-peng, LIU Er-gen, HU Xin-gen, WANG Guang-chao

(School of Basic Sciences, East China Jiaotong University, Nanchang 330013, China)

Abstract: By including degree, the conception of confidence degree of association rules are depicted and the relation between cinfidence degree and accuracy measure of Rough set are discussed. Especially, the incremental computing formulas of confidence degree are deduced by including degree.

Key words: including degree; confidence degree; accuracy measure.