相关测度与增量式支持度和信任度的计算*

王晓峰[1,2], 王天然[1]

[1](中国科学院 沈阳自动化研究所,辽宁 沈阳 110003);
[2](沈阳化工学院 计算机科学与应用系,辽宁 沈阳 110021)
E-mail: wangxf@mail.sy.ln.cn
http://www.sia.ac.cn

摘要: 通过相关测度的定义,从理论上探讨了增量式规则发现问题,并把分类规则挖掘和关联规则挖掘联系起来进行研究,为该问题的深入研究奠定了理论基础.相关测度刻画了给定关系和相关集合的数字特征.对相关测度的概念、定义、性质以及与支持度和信任度的关系等方面作了详细的分析和探讨,给出了基于相关集合的支持度和信任度的定义及计算方法.证明了测度增量定理和支持度增量定理,并给出了增量式支持度和信任度的计算公式.另外还详细地分析了数据增量对关联规则和信任度的影响,探讨了基于新支持度的候选项的修剪问题.所提出的相关测度及其思想为研究既能用于分类规则又能用于关联规则的统一数据挖掘方法提供了有价值的新思路.

关 键 词: 相关测度;支持度;信任度;关联规则;数据挖掘
中图法分类号: TP311 文献标识码: A

Agrawal 等人[1]在 1993 年提出了挖掘关联规则的一个重要方法——Apriori 算法.随后,有许多学者围绕提高算法的效率、减少数据库访问次数、增量计算等问题做了大量的研究工作[2~8].其中,文献[4]提出了分块挖掘大型数据库关联规则的方法,文献[5]提出了可用于数据有增减的关联规则挖掘算法,文献[6]提出了可变最小支持度增量算法,文献[7]提出了一个利用存储结构实现数据增量挖掘算法,文献[8]进一步把增量算法归纳为面向参数和面向数据两种类型,提出了序列模式增量挖掘算法的设计原则.这些研究工作表明,增量式关联规则挖掘方法越来越受到人们的重视.增量式挖掘方法有适应大规模、动态数据、降低内存需求、可实现并行处理等诸多好处,但增量式挖掘方法也有其困难之处,即由于把整个数据库的数据人为地划分开来,产生了分批挖掘出来的规则前后不兼容、规则遗失等问题.这些问题其实都与增量支持度和信任度的计算问题有关,现在对关联规则增量支持度和信任度的计算还是套用 Apriori 算法的定义,对支持度和信任度增量计算进行深入、细致的分析还不多见.本文试图从相关集合的角度出发,换一个思路来定义和计算支持度、信任度,利用相关测度定量地分析增量式支持度和信任度计算问题,以便较好地解决增量算法中规则的兼容性问题,从而进一步提高算法的效率.

1 相关测度的定义

在文献[9~11]中,我们定义了相关类、相关集合、相关强度等概念.对特定的关系、集合而言,相关强度是一个常量,但为了寻找最相关集合或扩张给定集合,需要一种动态的相关强度计算.因此,本文引入一个 $P(U)$到[0,1]上的映射 $F_A(X)$,称为相关测度.我们知道,若 X 是一个幂集的元素,那么它也是一个集合;如果 R 是定义在幂

* 收稿日期: 2001-02-06; 修改日期: 2001-04-18
基金项目: 辽宁省自然科学基金资助项目(9910200205);辽宁省教育厅高校科研基金资助项目(20012073)
作者简介: 王晓峰(1958－),男,辽宁灯塔人,在职博士生,教授,主要研究领域为数据挖掘,人工智能,数据库应用;王天然(1943－),男,黑龙江海伦人,研究员,博士生导师,主要研究领域为智能机器人.

© 中国科学院软件研究所 http://www.jos.org.cn

集上的一个二元关系,那么 R 描述的是集合间的关系.而集合之间的关系定义通常有两种形式,一种是用集合的内涵形式,如特征函数或语言描述;另一种是用外延形式,如集合的运算.相关测度反映了这种集合关系的外延数值特征,所以用内涵形式表达的集合关系要经过转化才能使用.我们称这种转化过程为"诱导产生".

1.1 相关测度定义

设 U 是非空有限集合,$P(U)$是 U 的幂集,R 是定义在 $P(U)$上的二元关系,若 $A\in P(U)$,给定集合 $X\subseteq U$,A 与 X 有关系 R,则称 $F_A(X):P(U)\to[0,1]$为集合 A 与集合 X 相对于关系 R 的相关测度.其中,$F_A(X)$是由关系 R 诱导产生的 $P(U)$到$[0,1]$上的映射,满足

$$F_A(\varnothing)=0;F_A(A)=1;0\le F_A(X)\le 1;F_A(X_i\cup X_j)\le F_A(X_i)+F_A(X_j).$$

这里,$X_i,X_j\subseteq U$,A 与 X_i,X_j 有关系 R.

注意:(1) R 是定义在幂集上的或集合间的二元关系.比如,A 与 X 有共同的属性关系,A 包含 X(包含关系)等等.通过关系,两个元素建立了一种联系(或关联),而相关测度是这种关联程度的一个量化.

(2) 相关测度描述了给定关系和相关集合的数字特征,但实际的 $F_A(X)$映射要根据关系的具体定义来考虑.如果关系是用某种集合运算定义的,那么只要它满足测度定义的条件就可作为 $F_A(X)$映射直接运用;如果是用语言或特征说明定义的,则必须把它用集合运算精确地描述出来,把关系化为集合间的运算形式.比如,设 $U=\{t_1,t_2,t_3,t_4,t_5\}$,$P(U)$是 U 的幂集,R 是 $P(U)$上的包含关系,R 可以精确地表示为 $R=\{\langle A,B\rangle|A,B\in P(U),A\supseteq B\}$,其中 A 的相关集合是$[A]_R=\{B|A\supseteq B,B\in P(U)\}$,这时 $F_A(X)=|[X]_R\cap[A]_R|/|[A]_R|$,满足测度定义的条件要求,是 R 的一种相关测度.

1.2 一个重要的相关测度

相关测度刻画了关系以及相关集合的数字特征,也是知识的一种数字化表征.特别是 $F_A(X)=|A\cap X|/|A|$,这是一个非常重要的相关测度,在数据挖掘方面具有重要的应用价值.为了方便应用,下面我们对它的运算性质进行必要的探讨.

性质 **1**. $F_A(A)=1,F_A(\varnothing)=0$,一般有 $0\le F_A(X)\le 1$.

证明略.

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

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

同理可证下面的性质 3~性质 10 成立.

性质 **3**. $F_{A_i\cap A_j}(X)=1-(1/|A_i\cap A_j|)(|(A_i\cap A_j)\cup X|-|X|)$.

性质 **4**. $F_{A_i\cap A_j}(X)=1-(1/|A_i\cup A_j|)(|A_i\cup A_j\cup X|-|X|)\le(1/|A_i\cup A_j|)(|A_i|F_{A_i}(X)+|A_j|F_{A_j}(X))\le F_{A_i}(X)+F_{A_j}(X)$.

性质 **5**. $F_{A_i\cap A_j}(X)=(1/|A_i\cap A_j|)(|A_i|F_{A_i}(X))+|A_j|F_{A_j}(X)-|A_i\cup A_j|F_{A_i\cap A_j}(X))$.

性质 **6**. $F_{A_i\cap A_j}(X)=(1/|A_i\cup A_j|)(|A_i|F_{A_i}(X)+|A_j|F_{A_j}(X)-|A_i\cap A_j|F_{A_i\cap A_j}(X))$.

性质 **7**. 如果集合 $X_1\subseteq X_2\subseteq U$,则 $F_A(X_1)\le F_A(X_2)$.

性质 **8**. 如果集合 $X_1,X_2\subseteq U$,则 $F_A(X_1\cap X_2)=F_A(X_1)+F_A(X_2)-F_A(X_1\cup X_2),F_A(X_1)+F_A(X_2)\ge F_A(X_1\cup X_2)\ge 0$.

性质 **9**. $F_A(X_1\cap X_2)=|X_1\cap X_2\cap A|/|A|=(|(X_1\cap X_2)\cap A|/|X_1\cap X_2|)(|X_1\cap X_2|/|A|)=F_{X_1\cap X_2}(A)(|X_1\cap X_2|/|A|)$.

性质 **10**. $F_A(X_1\cup X_2)=F_A(X_1)+F_A(X_2)-F_A(X_1\cap X_2)=F_A(X_1)+F_A(X_2)-F_{X_1\cap X_2}(A)(|X_1\cap X_2|/|A|)$.显然,如果 $X_1\cap X_2=\varnothing$,则 $F_A(X_1\cup X_2)=F_A(X_1)+F_A(X_2)$.

2 基于相关集合的支持度和信任度定义

2.1 Apriori算法中关于支持度和信任度的概念与定义

设 $I=\{i_1,i_2,\ldots,i_m\}$是商品标识码的集合,事务 T 是项目的集合,并且 $T\subseteq I.X\subset I,Y\subset I,X\cap Y=\varnothing$,则规则 $X\Rightarrow Y$ 的支持度(support)是事务集中包含 $X\cup Y$ 的事务数与所有事务数之比,即 $support(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|$.

如果事务集 D 中包含 X 的事务中有 c%的事务同时包含 Y,那么规则 $X\Rightarrow Y$ 在 D 中的信任度(confidence)是

© 中国科学院软件研究所 http://www.jos.org.cn

c,即 $confidence(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|X|$.

2.2 基于相关集合的支持度、信任度定义

假设 $D=\{t_1,t_2,\ldots,t_n\}$是一个事务集合,每个事务 t_i 有惟一的标识符 T_i 及若干种不同的商品项,那么把 D 中所有含商品 X 的事务(标识符)集中起来就构成了一个集合,我们称它为相关事务集,简称为相关集.为了方便,我们用 X 标识这个集合,$X=\{T_1,T_2,\ldots,T_m\}$,则$|X|$是含有商品 X 的事务数.

这样,若 X,Y 代表交易商品,那么 X,Y 同时也是事务集合.从交易商品方面来说,$X\in D,Y\in D$;从事务集合方面来说 $X\subseteq D,Y\subseteq D$.于是,新的支持度和信任度定义如下:

定义 **1**. 规则 $X\Rightarrow Y$ 在事务数据库 D 中的支持度是事务库中同时包含 X 和 Y 的事务数与所有事务数之比,即 $support(X\Rightarrow Y)=|X\cap Y|/|D|$.

定义 **2**. 规则 $X\Rightarrow Y$ 在事务数据库中的信任度是指包含 X 和 Y 的事务数与包含 X 的事务数之比,即 $confidence(X\Rightarrow Y)=|X\cap Y|/|X|$.

注意:(1) 这里,X 和 Y 是事务集合而不是商品集合,所以,一般 $X\cap Y\neq\varnothing$.

(2) X 和 Y 本身又是商品标识符,是一种或几种商品标识构成的符号串,X 和 Y 不重复.例如,$X=x_1x_2$ 代表标识符为 x_1,x_2 的两种商品,从集合上说是 X 又是 x_1,x_2 两个事务集的交集.

比较上述关于支持度和信任度的两种定义过程不难发现,虽然表面上看两种定义不一样,但从定义的含义(逻辑结果)来看,其本质是一样的.两种定义的主要差别在于支持度、信任度的计算过程,比如支持度计算,在前者包含 X 和 Y 的事务数的统计按集合的定义计算,在后者包含 X 和 Y 的事务数是以两个集合的交集形式给出来的,但最终结果完全一样.不难验证,信任度和支持度都是相关测度,只不过所涉及的关系运算有所不同.

2.3 信任度与Rough Sets精度的比较

当信任度的定义改成集合测度形式以后,把它和 Rough Sets[12]方法的精度(accuracy measure)相比较,不难发现两者是一致的.在 Rough Sets 中,给定等价关系 R 和非空集合 X,则关于 X 的精度为 $\mu_R(X)=|\cup\{Y\in U/R:Y\subseteq X\}|/|\cup\{Y\in U/R:Y\cap X\neq\varnothing\}|$.其中,$U/R$ 是 R 的等价类族;$|\cup\{Y\in U/R:Y\subseteq X\}|$表示所有属于 R 等价类又属于集合 X 的元素个数;$|\cup\{Y\in U/R:Y\cap X\neq\varnothing\}|$表示所有属于与集合 X 有交的等价类的元素个数.

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

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

比较相关测度的定义可知,在运算形式上两者完全相同,这说明粗糙度也是一种相关测度.

这不是巧合,而是有其自身的内在联系.因为 $confidence(X\Rightarrow Y)=|X\cap Y|/|X|$,$X,Y$ 是相关事务集,而每个事务是一个等价类,事务(项)集是等价类族,所以 X 是一个与 Y 相交的等价类的并集,而 Y 是一个待计算的集合.$X\Rightarrow Y$ 的信任度等价于由 X 计算 Y 的精确度.

另外,通过测度我们发现,包含关系、信任度和$\mu_R(X)$等表面上似乎完全不同的东西有着相同的测度计算公式,说明它们在数学本质上是相同的.因此,挖掘关联规则和分类规则没有本质上的区别,在相关测度计算上相同,说明两种挖掘方法能够融合在一起.这为我们进一步研究一种统一的挖掘方法奠定了基础,也为深入探讨数据挖掘的数学本质提供了一个新的思路.

3 支持度和信任度的增量计算

3.1 信任度的增量计算

信任测度增量定理. 设 U 是对象的有限集合,$F_A(X)=|A\cap X|/|A|$是集合 A 与集合 X 间的相关测度,$A\subseteq U,X\subseteq U$.数据库增加新数据记录以后,集合 $A'=A\cup\Delta A$,其中ΔA 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 $F_{A\cup\Delta A}(X\cup\Delta X)=\eta F_A(X)+(1-\eta)F_{\Delta A}(\Delta X)$.这里,$\eta=|A|/|A\cup\Delta A|$.

根据性质 10, $F_{A\cup\Delta A}(X\cup\Delta X)=F_{A\cup\Delta A}(X)+F_{A\cup\Delta A}(\Delta X)-F_{A\cup\Delta A}(X\cup\Delta X)$,注意到$|X\cap\Delta X|=|\varnothing|=0$,使 $F_A(X\cap$

© 中国科学院软件研究所 http://www.jos.org.cn

$\Delta X)=0$, $F_{A\cup\Delta A}(X\cup\Delta X)=0$,有 $F_{A\cup\Delta A}(X\cup\Delta X)=F_{A\cup\Delta A}(X)+F_{A\cup\Delta A}(\Delta X)$.

证明:由性质 4 有

$$F_{A\cup\Delta A}(X\cup\Delta X)=2-(1/|A\cup\Delta A|)[|A\cup\Delta A\cup X|-|X|+|A\cup\Delta A\cup\Delta X|-|\Delta X|]$$
$$=[|(A\cup\Delta A)\cap X|+|(A\cup\Delta A)\cap\Delta X|]/|A\cup\Delta A|.$$

注意,ΔA 对应的是新增数据记录部分,X 对应的是增量前数据记录,使得$\Delta A\cap X=\varnothing$,$|\Delta A\cap X|=0$;$\Delta X$ 对应新增数据记录部分,A 对应增量前数据记录,使得$|\Delta X\cap A|=0$.所以 $F_{A\cup\Delta A}(X\cup\Delta X)=[|A\cap X|+|\Delta A\cap\Delta X|]/|A\cup\Delta A|$.

令 $\eta=|A|/|A\cup\Delta A|$,注意到 $A\cap\Delta A=\varnothing$,$|A\cup\Delta A|=|A|+|\Delta A|$,有 $|\Delta A|/|A\cup\Delta A|=1-\eta$,$0\le\eta\le1$,得 $F_{A\cup\Delta A}(X\cup\Delta X)=\eta F_A(X)+(1-\eta)F_{\Delta A}(\Delta X)$. □

根据规则的信任度定义可知,$F_A(X)=confidence(A\Rightarrow X)$.所以本节测度的增量计算就是信任度的增量计算.因此有

推论 **1**. 若 $confidence(A\Rightarrow X)=1$,当 $confidence(\Delta A\Rightarrow\Delta X)=1$ 时,$confidence(A'\Rightarrow X')=1$.

这里,$confidence(A\Rightarrow X)$代表数据增量前规则 $A\Rightarrow X$ 的信任度,$confidence(\Delta A\Rightarrow\Delta X)$为增量部分规则 $A\Rightarrow X$ 的信任度,$confidence(A'\Rightarrow X')$为增量后规则 $A\Rightarrow X$ 的信任度(尽管符号不同,但表示的是同一条关联规则).证明从略.

3.2 支持度的增量计算

假设 X,Y 表示增量前的包含商品 X,Y 的事务集,$X\subseteq D$,$Y\subseteq D$,ΔX,ΔY 表示 X,Y 数据增量部分集合,ΔD 表示事务集 D 的数据增量部分,$\Delta X\subseteq\Delta D$,$\Delta Y\subseteq\Delta D$.令 $X'=X\cup\Delta X$,$Y'=Y\cup\Delta Y$,$D'=D\cup\Delta D$,则 X',Y'表示数据增量后的包含商品 X,Y 的事务集,D'表示数据增量后的事务集.有如下支持度增量定理:

支持度增量定理. 设 $S=support(X\Rightarrow Y)=|X\cap Y|/|D|$,$\Delta S=support(\Delta X\Rightarrow\Delta Y)=|\Delta X\cap\Delta Y|/|\Delta D|$,$S'=support(X'\Rightarrow Y')=|X'\cap Y'|/|D'|$,则 $S'=\lambda S+(1-\lambda)\Delta S$,其中 $\lambda=|D|/(|D|+|\Delta D|)$.

证明略.

推论 **2**. 如果 $|D|=|\Delta D|$,则 $S'=(S+\Delta S)/2$.

4 数据增量对关联规则兼容性的影响分析

4.1 支持度增量对频繁项的影响分析

设 $S=Support(X\Rightarrow Y)$是数据增量前项目集 XY 的支持度,$\Delta S=support(\Delta X\Rightarrow\Delta Y)$是数据增量部分 XY 的支持度,$S'=support(X'\Rightarrow Y')$是增量后 XY 的支持度.S^*是用户给定的最小支持度,概括数据增量前后项目集支持度变化情况对频繁项的影响,见表 1.

Table 1 The influence of incremental data on support of itemset

表 **1** 数据增量对项目集支持度的影响

S' \ S / ΔS	$Support(X\Rightarrow Y)\ge S^*$	$Support(X\Rightarrow Y)<S^*$
$Support(\Delta X\Rightarrow\Delta Y)\ge S^*$	$Support(X'\Rightarrow Y')\ge S^*$	$Support\ (X'\Rightarrow Y')=\lambda S+(1-\lambda)\Delta S$
$Support(\Delta X\Rightarrow\Delta Y)<S^*$	$Support(X'\Rightarrow Y')=\lambda S+(1-\lambda)\Delta S$	$Support(X'\Rightarrow Y')<S^*$

4.2 信任度增量对规则兼容性的影响分析

设 $C=Confidence(A\Rightarrow X)$是数据增量前规则 $A\Rightarrow X$ 的信任度,$\Delta C=Confidence(\Delta A\Rightarrow\Delta X)$是数据增量部分规则 $A\Rightarrow X$ 的信任度,$C'=Confidence(A'\Rightarrow X')$是增量后规则 $A\Rightarrow X$ 的支持度.C^*是用户给定的最小信任度,概括数据增量前后频繁项变化情况对频繁项信任度的影响,见表 2.

Table 2 The influence of incremental data on confidence of rule

表 **2** 数据增量对规则信任度的影响

C' \ C / ΔC	$Confidence(A\Rightarrow X)\ge C^*$	$Confidence(A\Rightarrow X)<C^*$
$Confidence(\Delta A\Rightarrow\Delta X)\ge C^*$	$Confidence(A'\Rightarrow X')\ge C^*$	$Confidence(A'\Rightarrow X')=\eta F_A(X)+(1-\eta)F_{\Delta A}(\Delta X)$
$Confidence(\Delta A\Rightarrow\Delta X)<C^*$	$Confidence(A'\Rightarrow X')=\eta F_A(X)+(1-\eta)F_{\Delta A}(\Delta X)$	$Confidence(A'\Rightarrow X')<C^*$

© 中国科学院软件研究所 http://www.jos.org.cn

4.3 支持度增量和信任度增量的递推计算

4.3.1 支持度增量的递推计算

假设增量算法将事务数据库中的数据划分成 N 个等量的数据集.设数据集为 $D_i,i=0,\ldots,N-1$,则支持度计算公式为

$$S_k'=\lambda_k S_k+(1-\lambda_k)\Delta S_k.$$

其中,$\lambda_k=|D_k|/(|D_k|+|\Delta D|)$,$|D_{k+1}|=|D_k|+|\Delta D|$,$S_k'$为第 k 次数据增量后事务集中项目 X,Y 的支持度,$S_k=S_{k-1}'$为第 k 次数据增量前事务集中项目 X,Y 的支持度,ΔS_k 为第 k 次数据增量部分事务集中项目 X,Y 的支持度.初始时令 $S_0=0,k=0,\ldots,N-1$;因为数据划分成 N 个等量的数据集,所以$\lambda_k=|D_k|/(|D_k|+|\Delta D|)=k/(k+1)$.

4.3.2 信任度增量的递推计算

假设增量算法将事务数据库中的数据划分成 N 个等量的数据集.设数据集为 $D_i,i=0,\ldots,N-1$,则信任度计算公式为

$$C_k'=\eta_k C_k+(1-\eta_k)\Delta C_k.$$

其中,$\eta_k=|A_k|/(|A_k|+|\Delta A_k|)$,$|A_k|$为数据集 $D_0\sim D_{k-1}$ 中含有项 A 的事务数之和,$|\Delta A_k|$为数据集 D_k 中包含项 A 的事务数.

C_k'为第 k 次数据增量后事务集中规则的信任度,$C_k=C_{k-1}'$为第 k 次数据增量前事务集中规则的信任度,ΔC_k 为第 k 次数据增量部分事务集中规则的信任度,初始时 $\eta_0=C_0=0,k=0,\ldots,N-1$.

注:(1) 若采用递推公式$|A_{k+1}|=|A_k|+|\Delta A_k|$,则 $\eta_k=|A_k|/|A_{k+1}|$;

(2) 若 $S_k(A)$表示第 k 次数据增量后事务集中项 A 的支持度,$\Delta S_k(A)$第 k 次数据增量部分事务集 D_k 中项 A 的支持度,则 $\eta_k=S_k(A)/(S_k(A)+\Delta S_k(A))$.

4.4 支持度增量计算和候选频繁项集的修剪

如前所述,Apriori 算法的成功之处在于,利用给定的最小支持度适时修剪不必要的项目组合分支,减少对无用候选项的计算.这种方法对于通常的一个事务数据集是有效的,但是对分块的增量数据集来说,会过早地修剪掉有用的候选项,出现规则丢失现象.为了解决这一问题,往往采用降低最小支持度的方法,但问题是不知支持度多大合适,太小会产生无用的项目计算,太大会丢失规则.

其实,这个问题本质上是一个误差舍取问题.由支持度增量式计算公式 $S_k'=\lambda_k S_k+(1-\lambda_k)\Delta S_k$ 可知,当支持度变化不大且增量次数较多时适时终止计算,会引起一个最大计算误差,这个误差可根据公式和计算次数估计出来.把每次计算的支持度和用户给定的最小支持度相比较,当递推公式的终止误差达到最大允许误差时结束计算,这样可以提高算法的效率,又不会丢失规则.如果设定支持度计算最大允许误差为 0,就可以保证不会丢失任何规则,否则,可以计算出规则丢失的最大几率.

5 结束语

本文利用相关集合重新定义了 Apriori 算法中的支持度和信任度,给出了数据增量时支持度和信任度的计算公式和递推算法,并讨论了提高增量算法效率的途径.所给出的计算公式、递推公式并没有具体指明用在什么样的数据挖掘算法中,实际上它们可以用于经典的关联规则算法,也可以用在粗集方法甚至另外开发的新算法中.我们把支持度和信任度统一归纳为相关测度,讨论了一个重要测度的定义、性质和增量定理.通过相关测度使粗集方法和 Apriori 算法的基本计算在数学上统一起来,为研究分类规则和关联规则的统一挖掘方法奠定了基础.这种统一的方法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,这也许是今后的一个重要的研究方向.另外,用本文提出的支持度、信任度概念和计算方法来设计或改进关联规则挖掘算法还有许多工作要做.

References:

[1] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S., eds. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. ACM Press, 1993. 207~216.

© 中国科学院软件研究所 http://www.jos.org.cn

[2] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In: Bocca, J.B., Jarke, M., Zaniolo, C., eds. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94). Santiago: Morgan Kaufmann, 1994. 487~499.

[3] Park, J.S., Chen, M.S., Yu, P.S. An effective hash based algorithm for mining association rules. In: Carey, M.J., Schneider, D.A., eds. Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. San Jose, 1995. 175~186.

[4] Savasere, A., Omiecinski, E., Navathe, S. An effective algorithm for mining association rules in large databases. In: Dayal, U., Gray, P.M.D., Nishio, Shojiro, eds. Proceedings of 21st International Conference on Very Large Data Bases (VLDB'95). Zurich: Morgan Kaufmann, 1995. 432~443.

[5] Cheung, D.W., Han, J., Ng, V., *et al*. Maintenance of discovered association rules in large databases: an incremental updating technique. In: Su, S.Y.W., ed. Proceedings of the 12th International Conference on Data Engineering. New Orleans: IEEE Computer Society, 1996. 106~114.

[6] Feng, Yu-cai, Feng, Jian-lin. Incremental updating algorithm for mining association rules. Journal of Software, 1998,9(4):301~306 (in Chinese).

[7] Yang, Xue-bing, Gao, Jun-bo, Cai, Qing-sheng. Incremental updating association rule mining algorithm. Mini Micro Systems, 2000,21(6):611~613 (in Chinese).

[8] Zhou, Bin, Wu, Quan-yuan, Gao, Hong-kui. Designing incremental mining algorithms of sequential patterns. Journal of Computer Research and Development, 2000,37(10):1160~1165 (in Chinese).

[9] Wang, Xiao-feng, Yin, Dan-na, Cheng, Shihchuan. Mutuality sets. Journal of Shenyang Institute of Chemical Technology, 1999,13(3):67~76.

[10] Wang, Xiao-feng, Yin, Dan-na, Cheng, Shihchuan. Mutuality sets and its applications in reduct of knowledge system. Journal of Tsinghua University, 1998,38(S2):6~9 (in Chinese).

[11] Wang, Xiao-feng, Tang, Zhong, Zhao, Yue. The application of mutuality sets for knowledge discovered from database. Journal of Nanjing University, 2000,36(11):52~57 (in Chinese).

[12] Pawlak, Z. Rough sets theoretical aspects of reasoning about data. Kluwer Academic Publishers, 1991.

附中文参考文献:

[6] 冯玉才,冯剑琳.关联规则的增量式更新算法.软件学报,1998,9(4):301~306.

[7] 杨学兵,高俊波,蔡庆生.可增量更新的关联规则挖掘算法.小型微型计算机系统,2000,21(6):611~613.

[8] 周斌,吴泉源,高洪奎.序列模式挖掘的增量式算法的设计原则.计算机研究与发展,2000,37(10):1160~1165.

[10] 王晓峰,尹丹娜,郑诗诠.相关集合及其在知识库化简中的应用.清华大学学报,1998,38(S2):6~9.

[11] 王晓峰,唐忠,赵越.相关集合数据库知识发现中的应用.南京大学学报,2000,36(11):52~57.

Correlativity Measure and Incremental Computation of Support and Confidence*

WANG Xiao-feng[1,2], WANG Tian-ran[1]

[1](*Shenyang Institute of Automation*, *The Chinese Academy of Sciences*, *Shenyang* 110003, *China*);

[2](*Department of Computer Engineering*, *Shenyang Institute of Chemical Technology*, *Shenyang* 110021, *China*)

E-mail: wangxf@mail.sy.ln.cn

http://www.sia.ac.cn

Abstract: By defining the correlativity measure, the problem of incremental discovering association rule is discussed in theory, and the mining association rule and the mining classification rule are combined to research, which establishes the theoretical foundations for researching the problem in detail. The correlativity measure depicts the numeral character of given relation and mutuality set. The conception, the definition and the properties of the proposed correlativity measure, and the relation between support and confidence are analyzed and discussed in detail. The new definition, methods of computing support, and the confidence based on mutuality set are proposed. The incremental computing formulas of support and confidence are given, and incremental theorems of

© 中国科学院软件研究所 http://www.jos.org.cn

support and confidence are also proved. On the side, the influences of incremental data upon association rules and the confidence are analyzed in detail. The problem of pruning candidate frequent item set based on new support is also discussed. The correlativity measure and its idea proposed in this paper provide a new valuable way for studying a unification method for mining classification rules and associate rules from database.

Key words: correlativity measure; support; confidence; associate rules; data mining

* Received February 6, 2001; accepted April 18, 2001

Supported by the Natural Science Foundation of Liaoning Province of China under Grant No.9910200205; the University Scientific Research Foundation of Educational Department of Liaoning Province of China under Grant No.20012073

www.jos.org.cn

www.jos.org.cn

© 中国科学院软件研究所 http://www.jos.org.cn