计算机科学2000Vol. 27№. 6

模糊关联规则的数据挖掘

Mining Fuzzy Association Rule

黄智兴　张为群

（西南师范大学计算机科学系　重庆 400715）

Abstract　In the real world, the relation between the data always represents fuzzy. In this paper, we introduce the fuzzy concept to the association rule data mining algorithm and give a set-oriented fuzzy association rule mining algorithm to enhance the function of the association rule.

Keywords　Fuzzy set, Support, Fuzzy Association Rule

1　引言

关联规则是指包含了一组对象间特定关联关系的规则。由于关联规则的挖掘有着广阔的应用背景，因此，人们对关联规则的挖掘算法作了大量的研究。从离散的布尔型变量、枚举型变量分析到连续的数值型变量分析；从平面的单事务项内部关系的分析到立体N维的多事务项之间关系的分析；从集中式的整体的静态分析到分布式的并行的分析和基于增量的动态分析。关联规则的挖掘算法不断完善，挖掘内容不断丰富，完整。然而，现实生活中，数据之间的关系通常表现为模糊关系，因此本文将模糊概念引入关联规则挖掘算法中，力图发现数据之间的模糊关系，使关联规则的挖掘功能得以进一步增强。

2　数值型关联规则的发现

2.1　布尔型关联规则和数值型关联规则

关联规则的发现任务是在给定的事务集中，找寻所有满足用户指定的最小支持度和最小自信度限制的关联规则[1]。如果物品在物品集中出现用'1'表示，否则用'0'表示，这种类型的关联规则的挖掘，我们称作布尔型关联规则挖掘；而在科学计算和商业领域的数据库中有着丰富的数据类型，这些类型可能是数值型（如工资、年龄）和枚举型（如邮政编码，职称等），关于这种数据类型的数据挖掘，我们称作数值型关联规则挖掘。发现数值型关联规则具体做法是：

1）将枚举型量按一定顺序映射到一个连续的正整数集上。

2）对数值型量采用两种办法：i）在事务数据库中数值型量的个数比较少的时候，将数值型量按其大小依次映射到一个连续的正整数集上；ii）在事务数据库中数值型量的个数比较多的时候，将该数值型量的取值区间划分成几个子区间，每个子区间对应一个正整数，事务数据库中的数值型量用相应的正整数代替。

这样数值型关联规则发现的问题转换成布尔型关联规则发现的问题，而布尔型关联规则发现问题可视作数值型关联规则问题的特殊情况。文[2]对如何划分子区间及规则合并等问题有详细的研究。

2.2　按区间划分的数值型关联规则存在的不足

1）在实际应用中，对于有些数值变量我们无法进行区间划分。比如针对股票上涨，下跌，持平这些模糊概念，如果将涨幅大于2%称为上涨，涨幅小于负2%称为跌，正负2%之间称为持平的话，那么涨幅为1.99%和涨幅为2.01%的两只股票将被划分到两个完全不同的区间；

2）对有些数值变量划分区间对我们无实际意义。如，在研究模糊神经网络的输入量与输出量的关系时，按照简单区间划分的方法则无法深入了解网络模糊规则的作用情况；

3）从已经发掘出的数值型规则中我们仅能从规则中得到条件发生时，结论成立的大致范围，而无法通过推理得到更精确的估计值。

以上不足，在我们引入模糊概念之后就将得到相应的解决。

3　模糊关联规则的定义

纵观各种形式的关联规则其本质是找寻一种集合到集合的映射，因此我们很容易联想到将集合的概念

黄智兴　硕士研究生。张为群　教授、系主任。

拓展到模糊集合，此时关联规则将拓展成模糊关联规则，下面给出模糊关联规则挖掘的相关定义：

定义1 设 $I=\{i_1,i_2,\cdots,i_m\}$，是一组属性集，$\tilde{P}=\{\tilde{p}_1,\tilde{p}_2,\cdots,\tilde{p}_n\}$是一组定义在属性集 I 上的模糊集的集合，m,n 为正整数。$I_V=I\times\tilde{P}$。如果模糊集 v 在属性 i 上有定义，则式子 $s(i,v)=t$ 表示属性 i 上的取值关于模糊集 v 的隶属度为 t，$(0\leqslant t\leqslant 1)$，这里，我们把隶属度也称作支持度(Support)，其中二元组 $\langle i,v\rangle\in I_V$。

定义2 D 是一组事务集(称之事务数据库)。其中每个事务 T 是属性集 I 上的一组取值，每个事务有一个唯一的标识符，称作 TID。

定义3 设模糊集的集合 $\tilde{X}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_k\}$，$1\leqslant k\leqslant n$，$\tilde{X}\subseteq\tilde{P}$，$k$ 为正整数，模糊集 $\tilde{x}_i$，$1\leqslant i\leqslant k$ 是定义在事务 T 的某一属性 j 上的模糊集，事务 T 在属性 j 上的取值关于模糊集 $\tilde{x}_i$ 的隶属度为 t_i，$0\leqslant t_i\leqslant 1$。如果 t_i 大于零，则称事务 T 支持模糊集 $\tilde{x}_i$。如果事务 T 支持 $\tilde{X}$ 中所有的模糊集，则称事务 T 支持 $\tilde{X}$，事务 T 对于 $\tilde{X}$ 的支持度为事务 T 对模糊集 $\tilde{x}_i$，$1\leqslant i\leqslant k$ 的隶属度 t_i 中的最小值，即 $\min t_i$，$(1\leqslant i\leqslant k)$，记为 $\text{support}(\tilde{X})$。

定义4 事务集对模糊集 $\tilde{X}$，$\tilde{X}\subseteq\tilde{P}$ 的支持度是指对事务集中所有事务支持模糊集 $\tilde{X}$ 的支持度的总和占事务数据库事务总数的百分比。

定义5 事务 T 对 $\tilde{X}\cup\tilde{Y}$，$\tilde{X}\subseteq\tilde{P}$，$\tilde{Y}\subseteq\tilde{P}$，$\tilde{X}\cap\tilde{Y}=\Phi$ 的支持度为 $\min(\text{support}(\tilde{X}),\text{support}(\tilde{Y}))$。如果事务集对 $\tilde{X}\cup\tilde{Y}$ 的支持度为 s，则称规则 $\tilde{X}\Rightarrow\tilde{Y}$ 在事务数据库 D 中具有 s 的支持度，记为 $\text{support}(\tilde{X}\Rightarrow\tilde{Y})=s$。

定义6 如果 $\text{support}(\tilde{X}\cup\tilde{Y})/\text{support}(\tilde{X})=c$，则称规则 $\tilde{X}\Rightarrow\tilde{Y}$ 的自信度为 c。记为 $\text{confidence}(\tilde{X}\Rightarrow\tilde{Y})=c$。

模糊关联规则的挖掘就是找寻在事务数据库 D 中满足用户指定的最小支持度和最小自信度限制的规则 $\tilde{X}\Rightarrow\tilde{Y}$。

通过上述定义，我们可以将布尔型关联规则和数值型关联规则的问题看成每个事务 T 对模糊集的隶属度要么为"1"或者要么为"0"的特殊情况，求事务集对模糊集 $\tilde{X}$ 的隶属度的总和就是对事务数据库 D 中支持模糊集的事务计数。这样，所有布尔型关联规则和数值型关联规则问题都可以转换成模糊关联规则问题进行处理。

4 模糊关联规则的数据挖掘的步骤

模糊关联规则的数据挖掘包括以下几个步骤：选择目标数据集，隶属函数的确立，事务数据库的建立，模糊关系的发现，模式解释与评价。其中，选择目标数据集的方法与一般的数据挖掘方法相同。

4.1 隶属函数的确立

模糊性是客观世界普遍存在的一种现象，隶属函数就是对这种模糊性的数学描述。在模糊关联规则的数据挖掘中可以通过不同的方法建立隶属函数，其中常用的有推理法、模糊统计法、择优比较法、绝对比较法、集值统计迭代法等[3]。模糊集隶属度的定义在整个模糊挖掘过程中是非常重要的。

4.2 事务数据库的建立

为方便我们对模糊关联规则的挖掘，不失一般性，事务数据库中的所用记录(Record)可以写成表(事务标识符 TID，属性 Attribute，模糊集号 FuzzySet，隶属度 Support)的形式。如果不存在定义在不同属性上的模糊集，事务数据库还可简写为表(事务标识符 TID，模糊集号 FuzzySet，隶属度 Support)的形式，这种格式更为简洁方便。我们把表中的一行称为一个事务项(Item)。

例：定义在属性 X 上有两个模糊集 A 和 B，A、B 的隶属函数分别为：

$$\mu_A(x)=\begin{cases}0, & x\leqslant 0\\ x, & 0<x\leqslant 1\\ 2-x, & 1<x\leqslant 2\\ 0, & 2<x\end{cases}$$

$$\mu_B(x)=\begin{cases}0, & x\leqslant 1\\ x-1, & 1<x\leqslant 2\\ 3-x, & 2<x\leqslant 3\\ 0, & 3<x\end{cases}$$

如表1中事务标识符 TID 为100，属性 X 的值为1.60时，通过映射在表2中产生两条事务项(100，A，0.40)，(100，B，0.60)，注：隶属度为零或特别小的事务项不加入新的事务数据库。

表1 映射前的事务数据库

TID	属性	值
100	X	1.60
110	X	2.40

映射 ⇒

表2 映射后的事务数据库

TID	模糊集	支持度
100	A	0.40
100	B	0.60
110	B	0.60

通过这样的方法，事务数据库中数值型的属性被映射到相应的模糊集上，并得出了对应模糊集的隶属度。

4.3 模糊关系的发现

目前人们对关联规则的挖掘算法已经作了大量的研究，如 AIS、SETM、APRIORI、DHP 等，其中面向集合(Set-Oriented)的方法进行关联规则的挖掘，可以充分利用集合操作的非过程性特点，算法可以用 SQL 语

句描述，使算法简洁易懂，而且在具体实现时又可以充分利用关系数据库中的各种技术来提高效率。此外，原有面向集合的挖掘算法也能较方便地过渡到模糊关联规则的挖掘算法，下面举例找出只含两个项目的模糊关联规则：

有数据插入表 Stock(tid，item，support)，其中 tid 表示交易号，item 代表与交易数据隶属的模糊集号，support 为交易数据关于模糊集 item 的隶属度。用 SQL 语句描述如下：

```
INSERT INTO C2
SELECT r1. item, r2. item, SUM(MIN(r1. support, r2. support))
FROM Stock r1, Stock r2
WHERE r1. tid=r2. tid AND r1. item<r2. item
GROUP BY r1 item, r2. item
HAVING SUM(MIN(r1. support, r2. support)) >=minsupport
```

我们在引入模糊概念后，挖掘算法并不比先前的复杂多少，同样还可以采用类似的优化技术对算法进行优化。模糊挖掘算法的优化的任务之一是减少候选集的个数，如定义在同一属性上的模糊集不能同时出现在一条频繁项目中，如果在连接表后产生的频繁项目集中包含上述项目(useless item)，应当从频繁项目集中删除。另一项重要的任务是减少连接表中的项目数，方法是删除掉那些在候选集中不出现的项目。例如算法 FSETM：

```
begin
    k=1,
    sort R1 on item;
    C1=generate supports from R1;
repeat
    k=k+1;
    if k=2 then
        begin
            sort R1 on trans_id ,item;
            filter R1 according to C1;
        end
    else
        sort Rk-1 on trans_id,item1,…,itemk-1;
        R'k=merge Rk-1,R1;
        sort R'k on item1,…,itemk;
        C'k=generate supports from R'k;
        C'k=remove useless items from C'k;
        Rk=filer R'k to retain supported itemset;
until Rk=Φ
end;
```

算法性能分析：模糊关联规则的发现算法是建立在面向集合的挖掘算法之上，不同之处是 SETM 算法是对满足条件的事务集计数，FSETM 算法是对事务集的支持度求和，而这并不增加算法本身的时间复杂度。并且 FSETM 还对 SETM 进行了优化，使得系统的 I/O 操作次数显著减少，具体的算法分析参见文[4]。

4.4 模式解释与评价

由于模糊集的隶属度定义的不同，可能对同一事务数据库挖掘的结果不同。因此，在对模式进行解释与评价的时候，应结合模糊集的隶属函数进行。同时，应对规则进行整理，去除无用的或冗余的模式，将有用的模式以适当的方式存储或返回给用户。

5 应用实例

在研究模糊关联规则的挖掘算法中我们对模糊神经网络中的 GARIC(generalized approximate reasoning-based intelligent control)模型[5]进行了实验，用该模型进行温度的模糊控制。由于事先对被控对象的性能无法完全了解，在控制过程中被控对象的性能也会因各种因素发生改变，我们在网络的自学习过程中需要了解控制规则在整个控制过程中的作用，以便调整控制规则达到更好的控制效果。

实验中 GARIC 网络的输入层有两个结点，对每个输入结点定义11个模糊集，模糊层有22个结点，模糊集的隶属函数均采用三角型隶属函数，规则层有121个结点，解模糊层有21个结点，即是说对输出结点定义了21个模糊集。图1为输入结点 e 的模糊集定义示意图，11个模糊集分别命名为 $E_1, E_2, \cdots, E_{11}$；类似地，我们将输入结点 $\dot{e}$ 的模糊集命名为 $F_1, F_2, \cdots, F_{11}$。输出结点 y 有21个模糊集，分别命名为 $T_1, T_2, \cdots, T_{21}$。

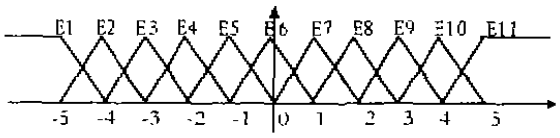

图1 输入结点 e 的模糊集定义

实验中记录总数为1500，在支持度为10%的情况下，发现下列规则(表3)

表3 模糊关联规则挖掘结果

e	ė	Y	支持度
E5	F4	T14	213.74
E5	F4	T15	178.54
E5	F5	T14	157.45
E5	F5	T15	272.52

利用数值型关联规则将 e，ė 的论域划分成11个区间，将 y 的论域划分成21个区间，在支持度为10%的情况下，仅发现一条规则，如表4。

表4 数值型关联规则挖掘结果

e	ė	Y	支持度
E5	F4	T14	344.00

(下转第63页)

络层的业务类型(ToS)优先级两种通用的类型，两者都是优先级方案，但均有其局限性。

区分服务体系是一种基于业务分类及其相关质量保证策略的体系，可认为是 ToS 的改进体系，但它是基于策略的服务质量机制。三者的比较见表2。

表2 三者比较

802.1P	ToS	DiffServ
基于优先级	基于优先级	基于策略/规则
在第二层数据包头增加16位	IP 包头 ToS 域 Precedence 域	IP 包头6位 DSCP 域
3位标记使用6种优先级	3位优先级位	64种类型
升级费用昂贵	7种优先级	基于 PHB 工作
与以前的网络不兼容		后向兼容 ToS
仅对第二层有效		

尽管 ToS 早在80年代就已定义，但直到最近为更好地保证不同的服务质量，才开始真正被应用在边缘路由器上。IPv4的包头中的 ToS 字节[4,5]如图2。

3位	优先级	级	D	T	R	C	0

图2 IP 包的 ToS 字节结构

ToS 字节包括 IP 包优先级子域和业务类型子域，在 RFC791中，3位优先级位提供7种分组相对优先级，这是一种业务相对优先级方案。优先级并不影响路由选择，而只影响分组队列，即当不同优先级的数据包在同一通道等待转发时，具有最高优先级的数据包最先被传发，同时这种优先级只在本网络内有效，优先级中最高的是网络控制数据分组，最低的是路由转发分组，剩余的5位没有利用。边缘路由器和符合 ToS 规范的第三层交换机根据分组的优先级位的值，将分组作相应的存储转发和丢弃处理。

在RFC1394中，用 ToS 字节的 DTRC 四位构成业务类型子域，来标识业务类型，对应路由选择方式。D、T、R、C 分别表示最小时延路由、最大吞吐量路由、最可靠路由和最小开销路由的服务质量需求。网络节点根据这些位的值选择相应的路由和转发方式。这是一种业务类型标记优先级方案。

DS 机制仍采用分类方法，与业务流相对优先级和业务优先级标记等方案不同的是，它基于网络管理策略和规则，不同的业务流被标以不同的转发方式 PHB，分配相应的网络资源。和7种 ToS 相对优先级和16种业务标记优先级相比，64种 DSCP 值的分配和使用规范及其对应的 PHB 更能满足网络业务的多样性，也更适应网络的管理和扩展需要。DS 机制应用现有的网络技术实现网络资源有效管理和更好的服务质量。

在目前 QoS 实现困难和成本昂贵的条件下，DS 机制通过更有效的网络资源管理策略缓减网络瓶颈，提高服务质量，减少网络设备的管理开销，也易于网络扩展。由于 DS 机制后向兼容 IPv4，对目前占领市场的流行 IPv4网络无需做大的硬件和软件改动，即可通过协议升级实现，从而保护了现有投资，较易令人们接受。可以这么认为，DS 是一种接近 QoS 服务质量体系但较 QoS 成本低廉的过渡服务质量机制。

参考文献

1 Blake S, et al. An Architecture of Differentiated Services. RFC 2475. Dec., 1998
2 Nichols K, et al Definition of the Differentiated Services filed (DS Filed) in the IPv4 and IPv6 Headers. RFC2474, Dec., 1998
3 Bernet Y, et al. Requirements of Diff-serv Boundary Routers Internet Draft. Nov., 1998
4 Information Sciences Institute. Internet Protocol RFC 791, Sept., 1981
5 Almquist P. Type of Service in the Internet Protocol Suite. RFC1349, July 1994
6 Guerin R, Peris V. Quality-of-Service in Packet Networks: basic mechanisms and directions Computer Networks, 31 1999, 31·169～189

(上接第42页)

利用模糊关联规则算法，发现了四条规则，我们清楚地看到 e、$\dot{e}$ 在落入 E5、F4、F5区间内时，GARIC 网络的输出情况，这对调整网络中的参数，调整网络控制效果，了解网络性能都很有参考价值。

同时，如果我们知道 e、$\dot{e}$ 在上述规则的作用域内，可以通过模糊推理的方法估计出 y 的值，在有充分的学习数据的时候，y 可以得出十分精确的估计值，而利用数值型关联规则无论有多少学习数据，始终只能得知有 y 的大致取值范围。显然利用模糊关联规则可以发现更为详尽的内容。

结论及展望 模糊的关联规则可以发现数据中存在的模糊关系，其结果使用户更容易理解，模糊关联规则可用于专家系统等应用中，并且在利用规则进行推理时，模糊关联规则将可以得出比其他数据挖掘的关联规则更详细更准确的答案。其发现内容还可以拓展到序列关系的发现、周期关系的发现；其算法还可以扩展到多概念层次的关系规则挖掘，事务数据库增量维护，分布式环境下关联规则采集等方面，将模糊数学与数据挖掘技术相结合是一项非常有意义的工作。

参考文献

1 史忠植. 高级人工智能. 科学出版社，1998. 206～208
2 Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Proc. of the ACM SIGMOD Conference on Management of Data, Montreal. Canada, June 1996 Available at: http: // www. almaden. ibm. com/cs/people/srikant/papers/sigmod96. pdf
3 汪培庄，李洪兴. 模糊系统理论与模糊计算机. 科学出版社 1996. 92～129
4 王寅北，夏庆，孙志辉. FSETM：一种面向集合关联规则的数据挖掘新算法. 见·第十五届全国数据库学术会议论文集. 1998. 144～147
5 张乃尧，阎平凡. 神经网络与模糊控制. 清华大学出版社，1998. 252～257