维普资讯 http://www.cqvip.com

多概念层次的数值关联规则挖掘

Quantitative Association Rules Mining with Multi-Level of Concept

高 飞 谢维信

(深圳大学信息工程学院 深圳518060)

Abstract Many progresses have been made on the studies of the algorithms of mining categorical association rules in the past years[1~4]. In practice, databases usually contain some quantitative attributes, but unfortunately algorithms of mining categorical association rules can not be applied directly to these databases. So it is necessary to provide the new definition of association rules in such situation. Based on the idea presented in[5], this paper presents the algorithm of quantitative association rules mining with multi-level of concept

Keywords Association rule, Categorical association rule, Quantitative association rule

1 引言

文[6]中将分类规则挖掘的方法扩展到数值关联规则挖掘的情况，其基本思想是：把一个数值属性 x 划分为若干个等分区间，于是一个三元组(x,l,u)便可对应于一个布尔项目，其中[l,u]表示一个数值区间，之后再用类似于布尔关联规则的挖掘算法进行挖掘。这种采用区间分割的方法来刻画数值属性，存在三点不足之处：1)致使信息丢失；2)经常使规则具有误导性；3)产生许多冗余规则。此外，通常某一数值属性的取值分布往往并非均匀的，因此等深分割(equi-depth partioning)就难以反映数值的真实分布及其对应的行为。文[7]中仍然沿用了文[6]中的定义，唯一不同的是采用了聚类方法对数值属性进行分割，这在一定程度上减少了规则的误导性。除了上述提到的关于数值关联规则的研究工作以外，还有一些学者的工作中也包含了数值关联规则的挖掘问题，如文[8]、[9]，但他们的工作往往偏重于预测，而并非发现关联规则。文[5]中基于统计学原理，给出了关联规则的一种新定义，但是该文中没有解决多概念层次的数值关联规则挖掘问题。本文对文[5]中的定义进行了适当的修改，并给出了多概念层次的数值关联规则挖掘算法。

2 与数值关联规则挖掘有关的若干概念

2.1 数值关联规则

关联规则反映了数据库中部分记录(或数据子集)所对应的一种有趣行为，例如：在一个零售商品交易数据库中，存在如下规则：啤酒⇒尿布，规则的左端，即啤酒，描述了数据库中购买啤酒的所有顾客，规则的右端定义了上述人群的一种行为，即购买尿布。因此，关联规则的一般结构可表示为：

数据集⇒有趣的行为

究竟何谓有趣的行为呢？在分类属性的情形里，有趣的行为意味着发生率较高的项集 Y，因此，行为自然用一个项集 Y 及其出现的频率(信任度)来描述，统计上来讲，信任度就是项集 Y 的条件概率分布($Pr(Y|X)$)。因此，一个分类关联可以表示成如下形式：

$X \Rightarrow Pr(Y|X) > min\_conf$

规则的左端表示记录集，而规则的右端表示了它的统计行为，当规则的右端为数值属性时，描述规则行为最好的方法显然是数值属性的统计分布，均值和方差是最常用的数值统计量。由于不论关联规则的右端为何种属性，均可解释为一种统计行为，因此数值关联规则可以看成是分类关联规则的一种扩展。

定义1 $I=(i_1,i_2,\cdots,i_n)$表示关系数据库 D 中的属性集，$I_c \subset I$ 是 I 的分类属性子集，$I_q \subset I$ 是 I 的数值属性子集，$I_c \cup I_q = I$，$I_c \cap I_q = \varnothing$。令 C_i 表示 I_c 的一个取值，$C=C_1,C_2,\cdots,C_m$ 表示 I_c 全部取值的集合。关系数据库中的一条记录可表示为：$t=((i_1,v_1),(i_2,v_2),\cdots,(I_n,v_n))$，$i_l$ 表示第 i 个属性，v_l 表示该属性的取值，v_l 为数值还是项目取决于属性 i_l 的类型。

定义2 项集 X 称为数据库 D 中的一个描述，如果满足 $X \in I_c \times C^T$，

定义3 一个项值关联规则具有如下形式：

$X \Rightarrow M_J(T_X) \quad (M_J(T_X) \neq M_J(D))$

此处，X 是一个描述，J 表示一个数值属性集，M 为一种统计度量，T_X 表示由 X 所定义的记录集(或数据集)，D 为数据库中全部的记录集，$M_J(T_X)$表示在记录

维普资讯 http://www.cqvip.com

集 Tx 上的关于属性集 J 的 M 度量，$M_i(D)$的含义与此类似。(注：≠符号表示统计不等，而非一般意义上的不等，后面我们将介绍如何利用假设检验的方法来判断两个统计量的不等关系。另外，本文只讨论 J 为单一数值属性的情况，关于多属性的情形，我们在后续文章中加以讨论。)下面是一些数值关联规则的例子：

1)(地区＝"辽宁")⇒平均工资＝250元/月(东北地区总平均工资＝150元/月)

2)(地区＝"沈阳")⇒平均工资＝250元/月(东北地区总平均工资＝150元/月)

3)(地区＝"辽宁"，行业＝"钢铁")平均工资＝250元/月(东北地区总平均工资＝150元/月)

4)(地区＝"辽宁"，行业＝"重工业")平均工资＝150元/月(东北地区总平均工资＝150元/月)

从上述四个数值关联规则的例子中可以看出，规则2～4所对应的数据集都是第一条规则所对应的数据集的子集，其中第2,3条规则是多余的，因为它们只是第一条规则的特殊情况，并且没有提供额外的信息，规则1、4则是有趣的，因为它们提供了新的信息，如第4条规则告诉我们，尽管辽宁省的工资在东北三省中明显较高，但是重工业工人的工资则处于一般水平。根据定义2，数值关联规则的比较部分通常是整个数据库，然而，比较一个数据子集与其所属的较大的数据子集往往也是有意义的，例如：将规则4的比较部分换成辽宁省的平均，则规则的含义变为：尽管辽宁省的工资明显高于东北地区的工资，但辽宁省重工业工人的工资则明显低于全省的平均水平。

2.2 规则及其子规则

定义4 有两个规则：1)$X_1 \Rightarrow M_i(Tx_1)$；2)$X_2 \Rightarrow M_i(Tx_2)$；如果满足：(1)$Tx_1 \subset Tx_2$；(2)$M_j(Tx_1) \neq M_j(Tx_2)$；则称规则1为规则2的子规则，规则2为规则1的父规则。如果又满足(1)式，则仍然称规则2为规则1的父规则，但是此时我们称规则1为规则2的准子规则。如果不存在任何规则位于规则1、2之间，满足该规则是1的父规则、2的子规则，则称规则1、2的父子关系为直接父子关系。

说明：一个子规则可以有属于它的子规则，于是不难发现规则与子规则间的这种层次关系。本文以下如不特殊说明，规则与子规则都是指具有直接父子关系的规则。利用这种层次关系，就可以有效地判断一个规则是否冗余规则：即它与其直接父规则是否满足统计不等。只有满足统计不等，它才能成为一个有效的规则。定义4是后面介绍的挖掘算法的主要依据。

3 数值关联规则的挖掘算法

3.1 频繁集及其层次关系

传统的关联规则首先要求 $Pr(X \cup Y) > min\_sup$，只有在此基础上才有必要判断 $X \Rightarrow Y$ 是否为一个符合条件的规则，即是否满足 $Pr(Y|X) > min\_conf$，如果满足，则它就是一个规则，否则就不是规则。在关系数据库条件下，根据定义2，知：规则 $X \Rightarrow M_i(Tx)(M_i(D))$是否为一个合格的规则，首先有满足 $Pr(X) > min\_sup$，在此基础上才有必要判断 $M_i(Tx)$与 $M_i(D)$是否统计不等，即 $M_i(Tx) \neq M_i(D)$，如果满足，则它是一个规则，否则就不是一个规则。在数值关联规则的定义里没有 min_conf 这一概念，而是利用了统计差异的概念，但两者在挖掘关联规则时所起的作用是一致的，它们都表示了 X 描述的数据集所对应的行为显著不同于比较项(或整体)所对应的行为。为了缩减规则的数量，我们约定：一个数值关联规则的比较项为其直接父规则所对应的数据集，不难发现，只有位于规则层次树的顶层的关联规则的比较项才为整个数据库 D，至于求解数值关联规则左端所对应的频繁集的方法参见文[10]中所介绍的 cumulate 算法。在求出频繁集之后，需要利用这些频繁集间的层次关系来进行数值关联规则的挖掘，后面我们将给出其挖掘的算法。图1是频繁集间层次关系的一个例子。

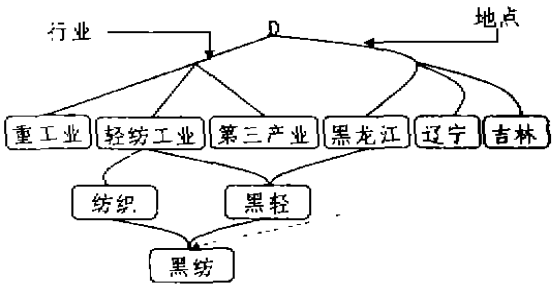

图1 频繁集间的层次关系

该数据库中的分类属性为行业与地点，其中行业在概念上存在特定的层次关系，如：纺织是轻纺工业的子节点。图中频繁集间的关系如下，第一排均为单项频繁集，其父集(或比较项)均为数据库 D，第二排中，单项频繁集纺织的父集为轻纺工业，而双项频繁集黑轻的父集为轻纺工业及黑龙江，最后，双项频繁集黑纺的父集为纺织和黑轻。由上图可以看出，在构造频繁集间的层次树时，频繁集与它的频繁子集间的构成关系：(1)由两个频繁集构成频繁子集，例如图中的由单个项目轻纺工业和黑龙江经过合并生成的双项频繁集黑轻；(2)利用项目的层次关系，对频繁集中的某个项目进行特殊化而生成频繁子集。如对频繁集黑轻中的"轻"进行特殊化，得到频繁集黑纺。有了频繁集的这种父子关系的层次结构，我们就可以有效地挖掘本文所

维普资讯 http://www.cqvip.com

定义的数值关联规则了。

3.2 数值关联规则的挖掘算法

算法：Find_Freq_Rule(comp_fi,fi)（从频繁集中挖掘数值关联规则的算法）。

输入：(1)关系数据库D中的全部频繁集及其层次关系；(2)门限 min_diff 及重要水平 α。

输出：数值关联规则及其子规则。

调用：Find_Freq_Rule(D),D)

方法：

```
Find_Freq_Rule(comp_fi,fi)
{
if(fi_dr≠Φ)do
 {
  for each x in fi_dr do
   {
    if fi=comp_fi)then x_dp=x_dp∪fi;
     else x_dp=x_dp∪comp_fi;
    end if
    //频繁集x所对应的规则只能是其直接父集所对应规则
      的子规则(定义4)
    if(∀ y∈x_dp),y_dp∉comp_fi 且 M_i(x)≠M(comp_fi))
then
      {
       Is_a_rule(x)=true;
       Output x as the sub_rule of comp_fi;
       Find_Freq_Rule(x,x);
      }
     else Is_a_rule(x)=false;
   }
  for each x in fi_dr
  if (Is_a_rule(x)==false)then Find_Freq_Rule(comp_
    fi,x);
  }
}
```

上述符号含义：min_diff：只有子集与父集的统计度量的差值大于 min_diff，才进行假设检验(统计差异的显著性检验)；α：假设检验的重要水平；fi：频繁项集；comp_fi：比较频繁项集；fi_dr：fi 的直接子集的集合(set of fi's direct children)；x_dp：频繁集 x 的直接父集的集合(set of x's direct parents)。

例：我们结合图1给出的频繁集说明算法的执行过程，首先调用 Find_Freq_rule(D,D)，找到D的直接子集轻纺工业与黑龙江，并得到如下规则：

轻纺工业⇒平均工资＝120元/月(东三省平均工资＝150元)

黑龙江⇒平均工资＝180元/月(东三省平均工资＝150元)

对于轻纺工业，我们调用 Find_Freq_Rule(轻纺工业,轻纺工业)；得到如下规则：

纺织⇒平均工资＝106元/月(轻纺工业平均工资＝120元)

黑轻⇒平均工资＝145元/月(轻纺工业平均工资＝120元)

对于纺织，再调用 Find_Freq_Rule(纺织,纺织)；得到如下规则：

黑纺⇒平均工资＝131元/月(纺织平均工资＝106元)

通过递归调用，该算法可以发现全部的数值关联规则。应该指出的是，尽管黑龙江也是黑纺的父集，但是它同时又是黑纺的直接父集黑轻的父集，因此黑纺不是它的子规则(定义4)。

3.3 统计不等的假设检验

由于在前述算法中采用了假设检验的方法来验证两个统计量的不等，因此这里我们对其进行一个简要的介绍。应用数据库中的数据分布很多时候是接近于正态分布的，如某一人群的工资分布，某一人群的月消费金额分布等等。如果用 A 表示某一规则所对应的数据集，B 表示 A 的直接子集，则可以把 B 看成是对 A 的抽样，于是可以利用样本 B 来估计总体 A 的统计特征。如果所估计出来的值显著不同于 A 的统计特征，则可认为 B 有明显不同于 A 的行为。常用的统计量是：均值、方差；在数据非常倾斜的情况下，往往采用分位数。为了比较两个统计量的差异，通常的方法是，先用样本数据构造一个统计量 z，再给出一个置信区间 $1-\alpha$，使得 $p(z\geq z_\alpha)<\alpha$（α 通常取 0.001、0.005、0.01 等），如果所计算的统计量 $z\geq z_\alpha$ 发生了，由于这是一个小概率事件，故而有两种可能性：(1)总体的分布不对；(2)所采集的样本有特异于总体的行为。由于总体的统计特征都是经过精确计算所得，故只能是后一种原因，即样本的行为显著不同于它所属的总体的特征。今后我们都假设关系数据库中的数值属性的取值为正态分布的。在正态总体的情况下，为了比较样本与总体的均值差异，可以构造如下的 Z 统计量：

$$Z=\frac{M(B)-M(A)}{\sigma(A)/\sqrt{N}}$$

其中 M(B)为样本均值，M(A)是总体的均值，σ(A)为总体的方差，N 为样本数。Z_α 可以通过反查 Z 分布表求得。Z 值告诉我们样本的均值与其所属的总体的均值相差有多远，例如 $Z_\alpha=2.57$，则样本位于正态分布99%以外的区域，也就是说从一次采样中获取如此高的 Z 值，其概率只有1%。同理，我们可以利用 X^2 统计量来检验样本的方差与总体的方差间的差别，关于假设检验的有关内容请参见文[11]和[12]。

结论 在对含有数值属性的关联规则进行挖掘时，最直接的方法是将数值属性分割为等间距的数值区间，以便可以将这些区间与分类属性同样对待，但是这种处理方法将带来信息的丢失和规则的冗余。本文所采用的数值关联规则的定义，更能体现数值属性的处理特点，即从统计学的角度对其加以处理。我们没有讨论数值属性出现在规则左端的情况以及多数值属性

(下转第55页)

维普资讯 http://www.cqvip.com

并触发SDH业务请求，该SDH业务请求在OTN层的配置处理流程与IP业务请求完全相同，但由于SDH业务请求是一个中介请求，它必须将OTN的资源分配结果反馈给原始的IP业务请求并由后者来验证本次业务配置是否成功。

3 IP/ATM

IP/ATM网络模式下的IP业务配置必须经历两次传递处理：即先经IP/ATM接口产生ATM业务请求，然后再经过ATM/SDH接口生成SDH业务请求，最后由OTN配置分配适当的网络资源。相应地，配置结果也要经历相反的传递才能反馈到IP层的业务请求。

与分层配置相比较，全光网络的聚合配置所带来的改善主要体现在以下几个方面：

(1)网络资源集中维护　由于业务层次仅生成业务请求，它们并不参与业务逻辑资源的分配，因此没有必要在这些层次设置逻辑资源库，业务请求的网络资源只有在OTN层才能得以调度和分配，这样，系统只需在OTN层设置网络资源库和调度策略库即可，从而保证了网络资源的集中维护。

(2)OTN层执行资源分配　在分层结构中，业务请求路径所通过的每一层次都参与资源的分配，这使得系统的处理相当复杂；聚合配置将业务请求直接或间接地传递到OTN层并由后者立即进行资源分配并反馈分配结果，它省去了不同层次间的资源映射过程，这种处理策略不但减少了层次间的关联特性(降低层次间的耦合度)，而且可以缩短业务的启动时间。

(3)兼容性强　OTN配置的业务请求适配模块支持各种业务接口的衔接，它不必关注业务层次的内部机制，因此，只要业务层能够向OTN层提供标准业务请求接口，OTN配置就能够满足相应的配置要求，这使得系统的兼容性大大加强。

(4)支持智能配置　智能化是未来网管的一个必然趋势，而聚合配置可以容易地对资源调度进行智能处理，图4中OTN层配置的策略库存储的主要内容就是资源调度算法，它使业务请求优先调度、资源使用时间限制、小业务复合等智能功能的实现变得容易实施。

小结　全光网络是一个一体化的综合宽带网络，它是多个业务网络和光传输网络的有机结合，如何有效地对它实施管理是业务正常供给的先决条件。本文针对全光网络的特殊性分析了传统分层配置管理和新提出的聚合配置管理之间的关联和差异，尽管聚合配置比较适合于全光网络的特征，但它的真正实现还必须依赖于业务层与OTN业务请求适配模块之间的接口标准定义，此外，调度策略库的知识构建也有待深入研究以便能够更好地提供智能化处理流程。

参考文献

1 王志文. 现代网络管理模型分析. 计算机工程与应用，1999，6
2 Doverspike R. Network Mangement Research in ATD-Net. IEEE Network
3 John Y. Connection Management for Multiwavelength Optical Networking. IEEE Communication
4 Lucent white paper. LUCENT TECHNOLOGIES OPTICAL NETWORKING 1999
5 Li Chung-Sheng, Ramaswami R. Automatic Fault Detection Isolation, and Recovery in Transparent All-Optical Networks. IEEE J. LIGHTWAVE TECHNOLOGY, 1997, 15: 1784～1793
6 Tanenbaum A. Computer Networks. 清华大学出版社，1997

(上接第87页)

的情况，这些工作有待于进一步的研究。

参考文献

1 Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proc. of the 20th Intl. Conf. on VLDB, 1994
2 Mannila H, Toivonen H, Verkamo A I. Efficient alogrithms for discovering association rules. KDD-94: AAAI Workshop on Knowledge Discovery in Databases, 1994. 181～192
3 Toivonen H. Sampling Large Databases for Association Rules. In: Proc. of the 22nd VLDB Conf. 1996
4 Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proc. 2000 ACM SIGMOD Int. Conf. Management of Data (SIGMOD'00), Dallas, TX, May 2000
5 Aumann Y, Lindell Y. A Statistical Theory for Quantitative Association Rules. In: Proc. of the 5th ACM-SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 1999
6 Srikant R, Agrawal R. Mining Quantitative Association Rules in Large Relational Tables. In: Proc. of the ACM SIGMOD Conf. on Management of Data. 1996
7 Zhang Z, Lu Y, Zhang B. An Effective Partitioning-Combining Algorithm for Discovering Quantitative Association Rules. In: Proc. of the First Pacific-Asia Conf. on Knowledge Discovery Data Mining, 1997
8 Fukuda T, Morimoto Y, Morishita S, Tokuyama T. Data Mining Using Two-Dimentional Optimized Association Rules: Scheme, Algorithms and Visulization. In: Proc. of the 1996 ACM SIGMOD Conf. on Management of Data, 1996
9 Yoda K, Fukuda T, Morimoto Y, Morishita S, Tokuyama T. Computing Optimized Rectilinear Regions for Association Rules. In: Proc. of KDD'97, August 1996
10 Srikant R, Agrawal R. Mining Generalized Association Rules. In: Proc. of the 21st VLDB Conf. Zurich, Swizerland, 1995
11 Larsen R J, Marx M L. An Introduction to Mathematical Statistics and Its Applications. Prentice Hall, 1986
12 周润兰，瑜胜华主编. 应用概率统计. 科学出版社，1999