移动通讯客户消费水平分析中的数据挖掘

■ 贺晋兵 博士生 刘云霞 博士生（厦门大学经济学院 福建厦门 361005）

▲ 国家社科基金资助(03BTJ014)，国家统计科学研究重点项目资助（LXZ0409）

◆ 中图分类号：F224.0 文献标识码：A

内容摘要：本文详细探讨了如何利用信息增益分析技术、属性的相关分析以及数据归约方法对数据库进行压缩，以及数据挖掘中关联规则、决策树和决策规则等方法的运用。并在此基础上，对我国某地区移动通讯用户消费水平的数据库进行了分析。

关键词：数据挖掘 事务项压缩 关联规则 决策规则

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。在数据挖掘过程中，最重要的步骤就是数据预处理，包含属性和元组归约的数据归约又是数据预处理中的关键环节。数据归约不仅压缩了数据库，也为决策规则和关联规则分析提供了前提条件。

数据归约技术

数据归约技术是将数据冗余压缩到最小，保证用尽可能少的有用信息进行数据的挖掘。下面将介绍几种数据归约技术，并将它们应用于对我国某地区手机用户消费水平的分析中。

（一）信息增益技术

信息增益是对属性包含信息量的度量，用信息熵表示。信息增益值越大说明某属性和其它属性差异越大，它的分辨能力越强，对分类的影响程度也就越大。设S是s个元组的集合，类属性中的分类有m个，设s_i是分别属于这m个类的样本数，$\frac{s_i}{s}$是S中样本属于该分类的概率估计，那么对于这个给定的样本分类信息熵是

$$I(s_1,s_2,\mathrm{K},s_m)=-\sum_{i=1}^{m}\frac{s_i}{s}\log_2\frac{s_i}{s} \qquad (1)$$

具有值$\{a_1,a_2,\mathrm{K},a_v\}$的属性A可以用来将S划分为子集$\{S_1,S_2,\mathrm{K},S_v\}$，其中，$S_j$包含中S中A值为$a_j$的那些样本，设$S_j$包含类$C_i$的$s_{ij}$个样本。则根据A划分的期望信息称作A的熵，它是加权的平均，即为

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\mathrm{K}+s_{mj}}{s}I(s_{1j},\mathrm{K},s_{mj}) \qquad (2)$$

根据A进行的划分获得的信息增益为

$$\mathrm{Gain}(A)=I(s_1,s_2\,\mathrm{K}\,s_m)-E(A) \qquad (3)$$

根据属性信息增益值的大小对属性重要性排序。设定“不重要”属性的阀值，删除信息增益值低于此阀值的属性。

下文对我国某地区移动手机用户2001年9月到2002年3月的缴费情况数据库的属性项进行第一步的压缩。该数据库的属性项有：号码、月份、用户类型、实际营收、月租、特服费、本地话费、长途话费、漫游费、信息费等10项，共计149632个事务项，我们称它为原始数据库T_1。

首先，按照“月份”将数据库T_1从2001年9月到2002年3月依次划分为7个子集$\{s_1,s_2,\mathrm{K},s_7\}$，根据式（1），这个样本分类的信息熵是$I(s_1,\mathrm{K},s_7)=2.80583$。

其次，计算每个属性的熵。如属性“用户类型”，它分为a、b、c(a为全球通用户，b为本地通用户，c为神州行用户)三种类型的熵分别：$I(s_{11},\mathrm{K},s_{71})=2.805763$、$I(s_{12},\mathrm{K},s_{72})=2.80641$和$I(s_{13},\mathrm{K},s_{73})=2.80487$；那么，如果样本按“用户类型”划分，则对给定的样本进行分类的期望信息是E(用户类型)=2.805433；这样，该划分的信息增益为Gain(用户类型)=0.000397。

同理，我们计算其它属性的信息增益分别是：“信息费”为0.007806、“实际营收”为0.162545、“月租”为0.010067、“特服费”为0.003164、“本地话费”为0.059884、“长途话费”为0.011242、“漫游费”为0.030309。设定阀值为0.01，删除属性“用户类型”、“信息费”、“特服费”。建立新的以“号码”、“月份”、“实际营收”、“月租”、“本地话费”、“长途话费”、“漫游费”为属性项的数据库T_2。

（二）属性的相关分析

用相关系数来描述属性项之间的相关程度，即

$$\rho_{AB}=\frac{E((X-E(X))(Y-E(Y)))}{\sqrt{D(X)}\sqrt{D(Y)}} \qquad (4)$$

如果相关系数小于0，则A的出现和B的出现是负相关的，一个值随另一个的减少而增加，这表明每一个属性都阻止另一个的出现。如果相关系数大于0，则A和B是正相关的，该值越大，意味着每一个的出现都蕴涵另一个的可能性越大。一个很大的相关系数表明A（或B）可以作为冗余而被去掉。如果值等于0，那么它们之间没有相关性。

利用相关分析对数据库继续进行属性的归约。根据式（4）计算T_2中各属性之间的相关性。“实际营收”和“本地话费”、“长途话费”、“漫游费”的相关程度都很高而与“月租”的相关程度并不高仅为0.3225，这是因为“实际营收”主要是由“本地话费”、“长途话费”和“漫游费”组成。“长途话费”与“漫游费”的相关程度是最高的，为0.8584；它们的变化趋势在相当高的程度上也相同，因此，可以将“漫游费”

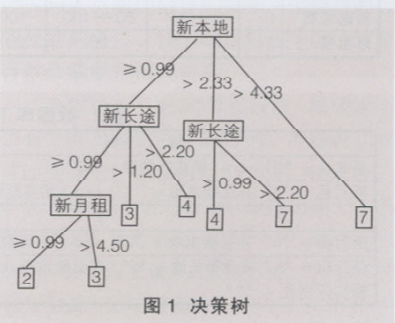

图1 决策树

作为冗余删去，建立新数据库 。

（三）进行数据库元组的压缩

数据库事务项的压缩，即是对连续属性进行离散化。在数据挖掘中属性的离散化有许多种方法，本文由于没有分类信息可以参考，并且考虑到话费的特殊性，笔者采用了“自然划分分段”的离散化方法，这样的离散化会使得数值区域被划分为相对一致、易于阅读、看上去较为“自然”。通过将属性项的域划分为区间，用区间标号来代替实际的数据值，就可以将连续数值离散化，离散化结果（见表 1）。

根据表 1 的赋值情况，合并数据库中属性值相同的元组，并建立新的数据库 。该数据库也是决策树和关联规则技术实施的基础。

决策规则与关联规则的挖掘

决策规则和关联规则是数据挖掘的两项主要技术，决策树是一个类似于流程图的树结构，利用一系列的规则划分，建立树状图，用于分类和预测。关联规则能够挖掘寻找给定数据集中项之间的有趣联系，这些规则能够找出客户的消费行为特点。

（一）决策树和决策规则

利用决策树和决策规则对数据库进行预测。将赋值后的属性项命名为“新实际营收”、“新月租”、“新本地话费”、“新长途话费”。以“新实际营收”作为分类属性，计算“新本地话费”、“新长途话费”以及“新月租费”的信息增益。将信息增益最高的属性项作为分区数据库的最初检验。依次类推，形成决策树（如图 1）。

从图 1 可看出，“本地话费”或“长途话费”只要大于 100 元，“实际营收”一般都会大于200元，这是一个较高的费用；而“月租”对于“实际营收”的影响不大。因此，移动运营商应该利用低月租及优惠的本地通话费策略以吸引更多的消费者来进行本地通话的消费。由此可将决策树应用在如何去确定一个先验信息的问题是很好的，使用决策树也可以得到有用的发现。决策树在变量的值能够相对地分成较少的不同数量时往往效果比较好。

（二）关联规则的挖掘

以“新实际营收”为目标变量，以“新本地话费”、“新长途话费”、“新月租”为输入变量，对数据库 T_4 进行多维关联规则的挖掘。同时满足最小支持度阀值(min_sup)为 0.08 和最小置信度阀值(min_conf)为 50%、最小频度为 5000 生成的规则为强规则。频繁项集分别为2、3、4，参见表 2。

利用关联规则我们可以发现很多有趣的或是相关的联系，如从表2关联规则的结果看，本地话费为 100 元以下的用户中有 20.79% 的用户，他们的消费水平是在 100 元到 200 元。而在全部用户中有 50.77% 的用户，他们的本地话费在 100 元以下同时消费水平在 100 元到 200 元之间。

决策树和关联规则结果的分析

从决策树和关联规则分析的结果可以发现：

高档消费的人数均是不同程度的在下降且幅度很大。1 档用户上升幅度很大，2 档的消费人数却是在下降，到3月份几乎各占到了一半。这主要是由于我国目前 IP 电话与传统长途电话的低廉资费使得一部分手机用户在长途电话方面尽量避免使用打长途，而选择使用 IP 电话或者固话，这是大部分客户长途话费处于低档水平的原因，也是长途话费与漫游费相关程度很高的原因。短信业务的发展也使一部分用户在处理非重要事情上选择了发短信进行交流。

本地话费以百元消费为主。每月的本地话费中，用户主要是集中在 2、3 档上。究其原因是通信手段的多样化，促进了有效竞争，手机用户有了更多的选择。无线市话通过低廉的资费赢得了大量用户，有的用户同时拥有小灵通和手机，分流了移动通讯的大量用户。

高档月租有固定消费群体。推出的 98 元、168 元、268 元、368 元、768 元的套餐业务有许多人选择。这部分消费者一般是企业骨干或是政府的职员，他们的工作大部分需要用手机联系来进行。他们在数据库中占有一定的比例并且消费相对稳定，实际营收也处于高水平，因此他们对于整个实际营收起着重要作用。

对原始数据库进行压缩也就是数据的预处理过程是数据挖掘的重要准备工作，而后再运用数据挖掘中的决策树和关联规则等技术进行分析，就可以将原始数据库中潜在的、重要的联系反映出来，从而帮助决策者进一步针对各种情况采取相应的对策提高收入和利润。数据挖掘技术还可以对大客户的消费习惯、偏好以及大客户长话主题进行分析，运营商便可以确定明确的客户关怀服务的目标客户群，这也是目前许多文献研究的主要内容。

表 1 各属性赋值表（单位：元）

赋值 / 项目	1	2	3	4	5	6	7	8
实际营收	0	0—100	100—200	200—300	300—400	400—600	600 以上	欠费
本地话费	0	0—100	100—200	200—400	400 以上			
长途话费	0	0—50	50—100	100 以上				
月租费	0	0—20	20	20—50	50	50 以上		

表 2 数据库 T_4 的多维关联规则表

规 则	置信度	支持度	n_频繁项
新本地 =“2”==>>新实际营收 =“3”	0.2079	0.5077	2
新月租 =“5”并且新本地 =“3”==>>新实际营收 =“4”	0.1204	0.5265	3
N	N	N	N
新月租 =“5”并且新本地 =“2”==>>新实际营收 =“3”	0.1727	0.6367	3
新月租 =“5”并且新长途 =“2”并且新本地 =“2”==>>新实际营收 =“3”	0.1064	0.8148	4

参考文献：

1.Paolo Giudici, David Heckerman, Joe Whittaker. Statistical Models for Data Mining[J]. Data Mining and Knowledge Discovery, 2001(5)

2.朱建平，张润楚.数据挖掘中事务性数据库的压缩及其应用[J].统计研究，2004(1)

作者简介：

贺晋兵，男，1979 年生，厦门大学经济学院计划统计系博士研究生，研究方向是统计学方法与数据分析。

刘云霞，女，1978 年生，厦门大学经济学院计划统计系博士研究生，研究方向是数据挖掘与数据分析。