第33卷 第24期 计 算 机 工 程 2007年12月
Vol.33 No.24 Computer Engineering December 2007
·软件技术与数据库· 文章编号：1000—3428(2007)24—0044—02 文献标识码：A 中图分类号：TP311

电信数据挖掘数据准备过程的规范化设计

蔡 鑫
（中国电信股份有限公司上海研究院信息集成部，上海 200122）

摘 要：从工程化实施电信数据挖掘项目的角度出发，在满足具体商业问题建模的数据要求前提下，对数据准备过程进行了结构化的分析和分解，提出一种规范化方法来约束宽表结构、源系统接口方式、数据预处理流程，并且预定义了相应的数据探索和数据准备过程，从源头改进电信数据挖掘项目的实施效率和质量。

关键词：数据挖掘；数据准备；宽表；规范化

Data Preparation Process Canonical Design for Telecom Data Mining

CAI Xin
（Shanghai Telecommunication Technology Research Institute, China Telecom. Co., Ltd., Shanghai 200122）

【Abstract】 Concerning telecom data mining project implementation and the data requirement of one business question, this paper analyzes the data preparation process, proposes a canonical plan for these items: wide-table structure design, source system data interface, data pretreatment process. It also pre-defines the of data exploration & data preparation to improve the of data mining project implementation efficiency and quality.

【Key words】 data mining; data preparation; wide-table; canonical

电信行业的竞争加剧，使得运营商更加关注管理活动的精确化，重视营销和服务的个性化。从本质上说，即更加依赖数据进行管理决策。随着技术手段的成熟，运营商开始尝试利用运营系统所积累的数据，采用数据挖掘的方法从中发掘出有价值的商业规则，建立了客户细分、流失预警、新业务响应等模型。经典的CRISP-DM数据挖掘方法[1]将数据挖掘项目实施过程分为商业理解、数据理解、数据准备、模型建立、模型评估、模型应用等阶段。在电信数据挖掘领域的研究中，比较多地关注于建模算法的选择和优化，这些经验对模型建立和模型评估有着很好的指导作用。但是，在实际的数据挖掘项目过程中，真正耗时费力的工作却是前期的数据准备过程，一般来说，需要占用项目50%~80%以上的时间，而当面对没有经过数据整合的多个孤立的BOSS运营系统时，问题将更加严重。如何利用规范化、工程化的方法，提高数据准备过程的效率，是数据挖掘项目实施者迫切面临的问题。

1 数据准备的复杂性分析

电信数据挖掘中数据准备过程的复杂性可以从以下方面分析：

(1)数据挖掘所要求的数据格式与传统IT系统中有较大差异。在普通的IT系统里，数据是以基本满足第三范式的实体关系模型存在的，这种结构有利于减少冗余、表达复杂的数据关系。而对于数据挖掘而言，大多数挖掘软件要求的输入形式则是扁平化的“宽表”格式。

(2)电信的IT系统众多，企业数据结构复杂。传统电信客户数据分布在计费账务、渠道支撑、呼叫中心、结算等多个系统当中，从如此多的分散的系统中抽取到需要的变量属性，并针对分析对象作整合和归并，其难度可想而知。对于已经建立了数据集成设施的数据环境，这种情形将有所改善。

(3)某些挖掘算法的特殊需要。例如：K_Means快速聚类，算法本身只是根据不同的样本点与聚类中心点的“距离”作为分群的依据，对于不同变量取值范围和量纲上的差异并不关注。如果直接将原始数据作为模型输入，建模的结果自然不会理想，需要根据具体变量的取值分布特点选择适用的标准化方法。

2 数据准备过程的任务

数据预处理方法有：数据清洗，数据集成，数据变换和数据归约[2]。

(1)数据清洗。以提高数据质量为目的，包括：数据的完整性及一致性检查，噪声数据处理，缺失数据填充，“脏”数据和重复记录消除等。

(2)数据集成。广义上是指将多个源系统的数据整合到一起，统一业务规则和编码规则，消除数据本身的冗余和冲突等；CRISP-DM中将数据集成（integratedata）定义为“对多个表格或者多条记录合并信息，从而建立新记录或新值”，是一种有所指的具体的操作。

(3)数据变换。将数据转换为适合挖掘的形式，可以根据需要构造出新的属性以帮助理解分析数据的特点，或者将数据标准化，使之落在一个特定的数据区间中。

(4)数据归约。是在尽可能保证数据完整性的基础上，获得数据的简化表示，以减少数据存储空间，使挖掘过程更有效，数据规约的概念很大，数据挖掘中常用和有效的是维归约，或称变量简约。

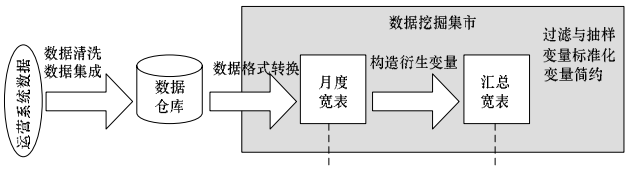
其中，数据清洗和跨系统的数据集成，是运营商企业构建数据仓库，力求解决的两件事情。在企业数据仓库的基础上实施数据挖掘项目，能够充分共享数据仓库的成果，将更适宜和更容易成功。从企业数据架构的整体布局来看，数据

作者简介：蔡 鑫(1975－)，男，工程师，主研方向：电信企业数据模型，数据仓库，数据分析

收稿日期：2006-12-30 **E-mail：**caixin@sttri.com.cn

挖掘的合理定位应该是数据仓库基础之上的、解决特定商业问题需求（报表、OLAP 所无法提供的潜在模式发现）的数据集市应用。

数据准备过程的总体流程如图 1 所示。



设备号码	姓名	年度月份	总费用	传统长话时长	总通话时长	……
28888888	张三	200601	80	1 000	20 000	
28888888	张三	200602	90	1 200	22 000	
28888888	张三	200603	85	1 100	18 000	
28888888	张三	200604	82	1 000	21 000	

设备号码	姓名	平均总费用	月平均传统长话时长	月平均总通话时长	传统长话通话占比	……
28888888	张三	84.25	1 075	20 250	0.053	
28888899	李四	68.60	980	10 000	0.098	

图 1　数据准备过程总体流程

3　数据准备的过程分解与规范化设计

根据以上原则和实践操作经验，将数据处理过程分为以下 6 个结构化的步骤：商业问题宽表模板设计，原始数据转换及月度宽表生成，衍生变量计算及汇总宽表生成，记录过滤和抽样，变量标准化处理，变量分析与简约。

3.1　商业问题宽表模板设计

所有的数据挖掘项目都是源于明确的商业问题，本文用问题类型和分析对象两个维度来描述一个商业问题：

(1)问题类型，如客户细分、流失预警、交叉销售等，决定了可以用哪些算法解决问题。

(2)分析对象，如政企客户、个人客户，或是固定电话用户、小灵通用户等，决定了宽表数据组织的大致结构。

确定了具体的商业问题后，可以用一种相对稳定的数据结构，支持对该商业问题的挖掘，这就为宽表模板的固化提供了依据。在宽表模板的设计中，需要综合考虑电信行业建立模型和解释模型所需的各类数据，主要包括：

客户背景数据（个人客户的年龄、性别、……，政企客户的规模、行业、……）

客户接触数据（设备拆装移、业务投诉、套餐更改、……）

客户价值数据（各业务消费额、异网通话占比、在网年龄、……）

客户行为数据（是否使用来电显示/彩铃/长途、使用时段、主被叫通话次数和时长、竞争对手业务使用情况、……）

为支撑数据挖掘的数据准备过程，设计了两种颗粒度的宽表格式：分析对象的月度宽表和汇总宽表（图 1）。其中，汇总宽表最终用于挖掘建模的数据结构，而月度宽表用作与源系统的接口表。

3.2　原始数据转换与月度宽表的生成

电信企业的 IT 系统众多，不同省市的系统在数据结构上差异很大，即使是已经实施了统一数据仓库的省份，物理模型也存在较大差别。因此，从现有的 IT 系统数据作为起点，来统一整个数据处理过程的处理逻辑是不可行的。

现实的做法是利用固化好的宽表格式作为与源系统的接口，从而在数据挖掘的后续流程中屏蔽源系统的结构差异，图 1 中的“月度宽表”就是用来完成数据接口功能的规范化格式。以这个规范的“月度宽表”格式作新的起点，就能实现数据处理后续过程的处理逻辑的固化和复用。

3.3　衍生变量的计算与汇总宽表的生成

月度宽表对于同一个分析对象，有多个月份的特征变量数据，为了适合挖掘算法输入格式要求，需要对这些变量进行进一步汇总和加工，生成更具一般统计意义的衍生变量，包括均值型变量、趋势型变量、占比型变量等。

均值型变量为对分析对象在观察月内的同一个特征变量取算术平均；趋势型变量对分析对象在观察月内的同一个特征变量计算相对增量；占比型变量，是对两个业务特征变量取比值，获得一个新的比例变量，以反映原变量之间的关系。

3.4　过滤和抽样的记录

过滤和抽样都是数据筛选的过程。其中，过滤主要指依据业务规则对数据记录的筛选，以满足商业问题的要求，比如在建立流失预警模型时，观察月期间内的新增用户和已流失用户通常在建模时会被过滤掉。

主要业务指标（如：ARPU/MOU）严重偏离总体分布的奇异值，也应当剥离出来单独进行分析。抽样主要从建模的技术需要进行考虑，合适的样本宜于模型训练，对于目标变量分布悬殊的预测类问题（如流失预警），为有效体现特征，一般需要重新调整各类的样本比例，即对分布处于劣势的一方进行过采样。

3.5　变量标准化处理

经过上面的过程，得到的变量根据实际意义有着不同的量纲和值域，对某些算法而言，如果直接将这些变量进入算法，难以获得理想的模型效果。取而代之的是要先将这些变量按照某种方法进行标准化（或称归一化）处理，使量纲和取值范围差异的影响得到消除。常用的标准化方法有：

(1)最小最大值标准化：把每个被标准化的项减去最小值，再除以极差。

$X' = (X - X_{\min})/(X_{\max} - X_{\min})$

易证：$-1 \le X' \le 1$

(2)Z 分数标准化：把数据标准化成一个均值为 0，标准差为 1 的 Z 分数。

$$X' = (X - \bar{X})/s,\quad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

可证：绝大多数的数据都会包含在 $[\bar{X} - 3s, \bar{X} + 3s]$ 区间内，因此，可以认为 $-3 \le X' \le 3$ 成立，对于范围以外的数据，宜作为孤立点剔除单独研究。

3.6　变量分析与简约

上面的变量，由于业务或其他原因，某些变量之间可能会有比较强的相关性，比如：传统长话月平均时长和传统长话月平均次数。在建模的过程中，同时使用这些变量是没有必要的，甚至对建模有副作用。变量简约的过程，就是以尽可能少的信息损失为原则得到一个适合的低维变量集。

主成分分析和因子分析，是统计学上常用的维简约方法，但结果产生的因子本身就是原有若干个变量的信息综合，即使经过因子旋转也难于解释[3]。因此，在实际操作过程中，使用 2 个较为简单和容易理解的步骤进行变量的简约：

(1)利用众数分析，找出“低效”变量

通过众数分析，找出在训练集中众数超过 95%的变量，特别是众数取值为 0 的变量，例如国际长途通话次数，除非要作该项业务的专题分析，这种变量对于建立一般模型功效不大，可以去掉。

(下转第 48 页)