数据预处理方法在移动通信行业中的应用

董 艳
(合肥工业大学 管理学院,安徽 合肥 230009)

摘 要:解决数据本身的质量问题,以某移动通信用户离网原因分析及预测为主题及为数据挖掘模型处理出需要的数据是文章的主要目的。文中运用了数据预处理中,维规约,属性集成与构造,多重插补,离散化,规范化,数据抽样等方法来得到一个完整的、近似真实的数据集。针对所处理数据含有大量缺失值的特点,选取了插补的方法进行处理。包括方法的插补方法的选择,到最后使用多重插补方法对缺失数据进行修正。预处理后的数据应用到具体数据挖掘模型后提高了数据挖掘的效率,降低了数据挖掘复杂度。

关键词:数据预处理; 数据挖掘;数据清洗;多重插补; 缺失值
中图分类号:TP274 **文献标识码**:A **文章编号**:1673-629X(2010)11-0225-04

Application of Data Pre-processing Method in Mobile Telecommunication Industry

DONG Yan
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: The main purpose of the article is solving the data quality, in order to pre-process data for data mining model customer churn analysis and prediction of a certain mobile telecommunication industry as a subject. The method used in the paper are: dimension reduction, integration and structural properties, multiple imputation, discretization, normalization, data sampling and other methods to get a complete, similar to the real data sets. In this paper, the processing of data containing a large number of missing values to the characteristics of the imputation methods for processing. It is including the method of imputation method of choice and using multiple imputation methods for missing data correction. The data after pre-processing applied to the data mining model improve the efficiency and reduce the complexity of data mining.

Key words: data pre-processing; data mining; data cleaning; multiple imputation; missing value

0 引 言

随着移动通信市场竞争的不断加剧,各大移动通信运营商都在不断改善自己的服务水平,加强管理能力,强化品牌意识。各大移动运营商都拥有自己独立的运营系统,且应用系统较完善,收集了大量数据,这为进行不同要求的分析提供充分的数据准备。然而原始数据庞大而冗繁,又由于数据库系统升级,早期入网要求不严格等各方面的原因,造成了数据在一定程度上的不一致和缺失。主要有以下这几种情况:

(1)早期入网用户不严格要求使用身份证,导致早期数据库中身份证信息包含较多脏数据。

(2)年龄等其他属性信息作为客户的重要特征信息,只能通过转换其它属性得到。

(3)由于各种数据挖掘和分析都是面向一定的主题,众多的数据表对数据挖掘主题并非全部有用,同时也加大数据挖掘工作的难度。

文中就是以移动通信用户离网原因分析及预测为主题的数据预处理过程。一般的数据预处理技术包括:数据清理(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)、数据规约(Data Reduction)[1~4]。数据预处理常常占据整个数据挖掘过程的30%到80%的时间[5],它的重要性正日益凸显。在具体的数据处理过程中,往往是不断反复的过程,要根据数据的实际情况进行调整。图1是文中数据处理过程的具体流程图。

文中根据实际需要先对数据库中重要数据表中的主要属性进行数据抽取和规约,在初步的数据集成和清洗后再进一步做数据变换和规约。每一个处理步骤

收稿日期:2010-02-03;**修回日期**:2010-05-22
基金项目:国家自然科学基金重点项目(70631003);教育部博士点基金(200803590007)
作者简介:董 艳(1982-),女,安徽合肥人,硕士生,研究方向为数据挖掘、数据库。

都包含了一定的数据变换，来适应下一步预处理过程的需求。文中在说明如何进行数据预处理的同时，重点介绍了在含有不一致和缺失数据情况下采取多重插补的方法进行数据修复，从而构造出完整的样本数据。以下将对具体处理步骤展开介绍。

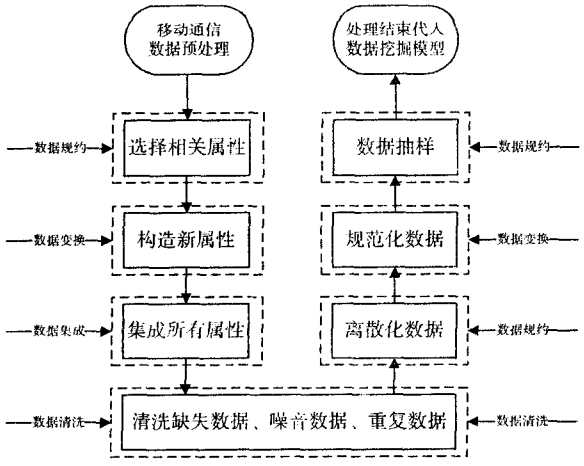

图 1 数据预处理流程图

1 数据规约

1.1 维规约

移动通信数据库中的数据是面向应用的，进行的数据分析是面向主题的。我们分析的数据库中拥有七十多张数据表，而文中面对的数据挖掘主题需要两类数据：一类是入网时用户提供的用户数据；另一类是用户入网后的历史交易数据。通过使用贪心算法[6]中的逐步向前选择技术对数据进行了筛选和分析。从数据库中找出了 5 个与分析主题有关，而且可以满足分析需要的数据表，并从中提取出有用的属性：

(1)客户资料表。存放所有客户的基本信息，包括 28 个数据项，其中与分析主题有关的有 5 个数据项：客户 ID，客户名称，客户建立时间，信用度，证件号码。

(2)用户资料表和离网用户资料表。存放用户的基本信息，包括 26 个数据项，其中与分析主题有关的有 5 个数据项：用户 ID，客户 ID，品牌代码，状态代码，状态变化时间。

(3)历史欠费表。存放用户的欠费信息，包括 31 个数据项，其中与分析主题有关的有 4 个数据项：用户 ID，客户 ID，帐务年月，品牌代码。

(4)综合帐单表。存放用户的每月总费用基本信息，包括 29 个数据项，其中与分析主题有关的有 6 个数据项：用户 ID，客户 ID，帐务年月，品牌代码，应收费用，优惠费用。

(5)明细帐单信息表。存放用户的每月费用详细信息，包括 16 个数据项，其中与分析主题有关的有 8 个数据项：用户 ID，帐目，客户 ID，帐务年月，应收费用，优惠费用，拨打次数，通话时长。

1.2 属性的集成与构造

以上对分析主题所需要使用的属性进行纵向选取，现在将相关属性进行集成。通过客户 ID 将客户资料表、用户资料表合并，生成包含 10 个属性的临时表 SAMPLE_01(见表 1)。

表 1 SAMPLE_01

客户 ID	用户 ID	客户类型	性别	年龄
状态代码	状态变化时间	客户建立时间	信用度	品牌代码

其中年龄、性别均由身份证信息中提取。在原始表中日期的格式并不完全一致，所以在合并时将所有日期格式统一为年月的形式。SAMPLE_01 中所含的数据大部分是通过用户告知所得，故有很多属性的数据出现缺失和噪声，将在后面的工作中进一步进行清洗。表 2 为临时表 SAMPLE_02。

表 2 SAMPLE_02

用户 ID	客户 ID	帐务年月	品牌代码	月总消费额	本地通话费
拨打客服电话次数	短信费用	开通增值业务数量	GRPS 费用	漫游费用	长途费用

通过历史欠费表和离网用户资料表对离网用户进行统计和转换，生成了临时表 SAMPLE_03，共计有 6 个属性，分别是：用户 ID，客户 ID，品牌代码，离网方式，离网时间，欠费次数。其中主要的属性是用户离网的具体日期，用户的所有欠费次数是通过统计的方法计算得来。

以上三个表的处理方法都是在 ORACLE 数据库中运用 SQL 语言的各种函数和过程实现的。至此将原库中所能提供的对分析主题有用的原始有用信息都包含在内了。但是这些信息依然过于粗糙，需要进行进一步的清洗和处理。

2 数据的清洗和规约

在真实的数据库中大量存在缺失数据、异常值和重复数据。其中缺失数据产生的原因分为两类：一类是机械原因，一类是人为原因。文中涉及的缺失数据主要指由人为原因造成的缺失。在原始表中的缺失值多数为由于输入错误或者是客户没有提供真实有效的信息，造成在统计分析时发现有很多异常值。例如：年龄为 123，-12，3；普通用户的姓名多于 4 个字符或非汉字字符，性别出现 0 或 1 以外的其它值。这些通过客户自行填写或由工作人员手动输入的数据，都可能存在异常值。将这些异常值进行了清理，使得这些异常值成为缺失数据的一部分。

2.1　异常数据和缺失数据的处理

原始数据中有很多重要属性的数据有缺失,如果只是简单的抛弃缺失值,对数据挖掘的效果存在一定影响,也改变了原始数据的真实有效性。有些属性对分析主题至关重要,不可以随便舍弃。但是大量的数据缺失给分析工作带来了困扰。为此,采取了对数据进行插补的方法,将重要属性进行插补。下面将对插补的方法做简单介绍。

2.1.1　插补方法的选择

在处理缺失数据时多数采取以下方法进行处理:(1)删除含有缺失值的个案;(2)可能值插补缺失值。其中插补方法有:(a)均值插补;(b)多重插补等。以上两种插补方法,均值插补是早期常用插补方法。插补过程比较简单,易于实现。由于是单值插补,仅适合缺失数据比较少,原始数据量小的情况下使用,如果缺失率过高就会造成插补效果不理想。多重插补,是面向大数据集填充方法的典型代表[7]。多重插补(MI)是Rubin(1978 年)[8]首先建议的,此类算法一经提出就受到了广泛的推崇,并已被应用到很多领域。目前许多软件公司已经集成了多重插补的算法到软件中,例如,SAS 软件在文献[9]中就将 MI 作为正式的过程加以使用。这对多重插补的进一步推广起到了推波助澜的作用。当然它的运用也是有条件的,在数据缺失率小于 15%[9]时与普通的插补方法没有太大的差别。然而随着样本量的增加或数据缺失率上升到 15% 至 35%[9]时,多重插补显示出了明显的优势。插补的效果和精度明显高于均值插补。由于现在硬件价格的下降,可以通过提高插补次数提升插补的效果,使得到的插补数据更加接近真实值。但是当缺失率高于 35% 时这几种插补方法都失去了有效性[9]。文中涉及数据的最高数据缺失率为 29.7%,且缺失数据属于随机缺失(MAR)类型[8],即观测缺失值的概率只依赖于观测值本身,并不依赖于缺失值,符合应用多重插补的条件。故文中将采取多重插补的方法进行缺失数据的插补。

2.1.2　多重插补的原理和应用

多重插补(MI)[10]是用一个 $D \geqslant 2$ 的插补向量代替每一个缺失值的程序。D 个值是有序的,即由插补值的向量可构造 D 个完全数据集,每一个缺失值用它的插补向量的第一个分量代替,构造出第一个完全数据集,每一个缺失值用它的插补向量的第二个分量代替,构造出第二个完全数据集,如此继续[8]。每个插补数据集合都用针对完整数据集的统计方法进行统计分析,对来自各个插补数据集的结果,根据评分函数进行选择,产生最终的插补值。数据缺失率在 15% ~30% 时,经 MI 处理后的分析结果更接近'真实'[9]。文中利用 SAS 中的多重插补过程对年龄、性别等重要属性分别进行了不同样本集的多重插补。通过上文中所采用的各种预处理方法所得到的数据样本依然庞大,有 683715 条记录,其中 age 的空值为 193311 个,缺失率为 28.3%,满足条件需求。这样的样本集加大了插补的难度和后期数据挖掘计算的复杂度。故采取样本抽样技术中的等距抽样技术对源数据进行了样本抽样。取原样本的 1/6.8,得训练样本 100547 个。其中 age 空值率为 28.3%。表 3 是对等距采样情况下年龄属性的两种插补方法的比较。

表 3　100547 个样本下插补方法比较

年 龄	多重插补	均值插补	原始数据
<16 岁	862	283	283
16 - 25 岁	34778	26718	26718
26 - 40 岁	34611	52533	24049
40 - 50 岁	20816	13859	13859
50 - 60 岁	7209	5158	5158
>60 岁	2271	1996	1996

从插补结果来看虽然均值插补和多重插补都将缺失数据补全,但是均值插补的峰值变化趋势已经远离了原始数据。而多重插补与原始数据的分布有所不同,但是在总体趋势上大体保持了一致,可见多重插补方法明显优于均值插补。故将多重插补后的结果数据作为数据挖掘模型的输入数据。

2.2　重复数据的处理

文中所处理的原数据问题属于单源数据质量问题,所以在异构数据库中的一些命名冲突、结构冲突、重复记录等问题并未出现。故文中处理的重复数据问题[11]与数据本身意义相关。例如一个客户同时拥有多个号码,将此用户的多个号码的消费额汇总求均值来代表此用户的真实消费额。客户只要有一个号码在使用,就不算是离网用户等。在进行各表汇总时会出现重复记录,也要通过 SQL 语言中的 distinct 等方法进行唯一化处理。

2.3　数据集成和离散化

在上面的处理过程结束后将得到初步的汇总表 SAMPLE_04,其中包括 18 个属性。

表 4　SAMPLE_04

客户 ID	客户类型	性别	品牌代码	信用度	状态代码
拨打客服电话次数	短信费用	年龄	当月总消费额类型	欠费次数	本地通话费
开通增值业务数量	GPRS 费用	漫游费用	长途费用	离网方式	离网时间

选出来的这些属性,现在已经能够初步满足分析的需要。但是为了加快数据挖掘模型的学习速度,对

训练样本集进行了规格化。主要是对相应属性进行离散化处理。规范化方法有很多种,文中主要对年龄,每月总消费额,信用额度进行了离散化操作。将年龄字段根据总的消费特征,离散化为离散值 1,2,3,4,5,6。它们分别代表 16 岁以下,16 到 25 岁,25 到 30 岁,30 到 40 岁,40 到 50 岁,50 岁以上的人群。将用户每月消费额和信用额度也做类似处理。

至此所有单月数据处理的初步工作就全部完成了,其它月份按照类似步骤处理即可。在代入模型前可根据模型需要抽样出不同数据量的样本训练集和属性,文中就是代入 COX 模型[12]进行数据挖掘。

3 结束语

在实验初期,并没有对原始数据进行以上处理,只能选取消费总额,信用额度,是否欠费,是否离网等原始信息作为输入数据,但是这些数据包含的信息量过少,数据噪音很大。无论取多少个训练样本,也无论如何调整模型的参数,分析结果都非常的差,无法进行下一步的数据挖掘工作。通过采取以上的预处理得到的训练样本,属性相关性,数据规范性都得到了很大的提高。

不能直接对原始数据进行分析的原因可能有以下几种:

(1)输入的原始属性虽多,但缺少通过转换得来的与分析主题真正相关的属性,偏离了分析主题;

(2)原始数据含有噪声,缺失,重复等各类数据问题,这对各类数据挖掘模型的应用带来了干扰;

(3)没有对数据进行规格化处理以及用科学的方法进行采样,很多数据不符合数据挖掘模型的要求。

文中就移动通信数据中离网用户数据的数据挖掘中的预处理过程进行了讨论,详细描述了数据抽取,数据清洗,数据转换的全过程。尤其重点阐述了关于缺失数据处理的过程,通过全面分析数据特征和插补方法,最终选定使用多重插补方法对数据进行修正,突破了一直以来采用的简单丢弃和插补单一值的方法,使修正后的数据更加接近真实值。此类处理数据的方法对处理通信业以及其它拥有海量数据的行业,进行数据预处理具有一定的借鉴意义。

众所周知,数据挖掘是个反复的过程,与此同时数据预处理也是一个不断优化的过程,要根据分析主题的不同进行相应的处理,往往最初几次的处理效果都很难理想。如何能利用更短的时间处理出更加有用的数据,还需要不断的努力探索。

参考文献:

[1] 安淑芝.数据仓库与数据挖掘[M].北京:清华大学出版社,2005.

[2] Han Jiawei,Kamber M. 数据挖掘概念与技术[M].北京:机械工业出版社,2001.

[3] 李雄飞,李 军.数据挖掘与知识发现[M].北京:高等教育出版社,2003.

[4] Rahm E,Do Hong hai. Data Cleaning:Problems and Current Approaches[J]. IEEE Data Engineering Bulletin,2000,23(4):3-13.

[5] Dasu T,Johnson T. Exploratory Data Mining and Data Cleaning[M]. USA:John Wiley & Sons,Inc. Publication,2003.

[6] 王洪涛.数据预处理技术的研究与实现[D].沈阳:东北大学计算机学院,2002.

[7] 胡红晓,谢 佳,韩冰. 缺失值处理方法比较研究[J].商场现代化,2007,15:352-353.

[8] Litle R J A,Rubin D B. Statistical Analysis With Missing Data[M].孙 山译.北京:中国统计出版社,2004:10-74.

[9] 殷 杰,石 锐.SAS 中处理数据集缺失值方法的对比研究[J].计算机应用,2007,27(6):438-439.

[10] 杨 军,赵 宇,丁文兴. 抽样调查中缺失数据的插补方法[J].数理统计与管理,2008,27(5):821-831.

[11] Muller H,Freytag J C. Problems, Methods, and Challenges in Comprehensive Data Cleansing[EB/OL]. 2003. http://www.dbis,informatik.hu-berlin.de/fileadmin/reserach/papers/techreports/2003-hub-ib-164-muller.pdf.

[12] 卢纹岱. SPSS for Windows 统计分析[M].第 3 版 .北京:电子工业出版社,2006:571-578.

(上接第 224 页)

[8] 李 波,石冰心,沈 斌,等.网络资源管理和仿真工具研究进展[J].微机与应用,2005,24(3):4-7.

[9] Fan R,Cheded L, Toker O. Internet-based SCADA: A new approach using JAVA and XML[J]. Computer putting & control engineering,2005,16(5):22-26.

[10] Viswanath P, Tse D. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality[J]. IEEE Trans. Inf. Theory, 2005,51(2):506-522.

[11] Hui S C, Leung M K H. Eleview :remote intelligent elevator monitoring system[J]. International Journal of Computers and Applications,2004,262(2):111-118.

[12] 伍云霞,孙继平.智能监控分站数据采集与监控程序实现方法[J].煤矿机电,2005(4):67-68.