

数据挖掘技术在客户关系管理系统中的应用

Application of Data Mining Technology in the CRM System

吴晓渊¹ 孟兆炜²

Wu Xiaoyuan Meng Zhaowei

(1. 桂林陆军学院计算机教研室 桂林 541002;

2. 国防科技大学计算机学院 长沙 410073)

(1. Comp. Staff Room, Guilin Military Academy, Guilin, 541002;

2. Coll. of Comp., National Univ. of Defense Tech., Changsha, 410073)

摘要 阐述数据挖掘的概念和主要分析方法, 以及数据挖掘技术在客户关系管理系统中的应用, 并结合一个客户流失分析的实例说明用于客户关系管理系统的数据挖掘工具的设计和实现。

关键词 数据挖掘 客户关系管理 客户流失分析

中图法分类号 TP311.131

Abstract The conception and the major analytical method of data mining, the application of data mining in the Customer Relationship Management (CRM) are introduced, and the design and implement of data mining technology in CRM is illustrated by an example of churn analysis.

Key words data mining, CRM, churn analysis

基于 Internet 技术的电子商务, 正在不断改变着企业的经营模式和组织架构。在以客户为中心的时代, 借助信息技术发现潜在的新客户以及保持并改善与客户的关系已成为企业的迫切需求。所以, 近年来客户关系管理系统得到了迅猛发展。客户关系管理 (Customer Relationship Management, CRM) 是一个全面体现以客户为中心思想的电子商务平台, 它应用到企业与客户相关的市场、销售、服务、技术支持等各个部门, 整合企业所有与客户相关的资源, 帮助企业改善与客户的关系。

客户关系管理的核心观念是以客户为中心, 即充分理解客户的需要, 为客户提供个性化的与其需求相适应的产品和服务。但是, 随着企业的成长和信息的累积, 很多企业逐渐发现要从其累积的海量客户信息和交易数据中发现客户需求规律已经变成极为困难的事情。因而, 能够从海量数据中挖掘有用的规律和知识的数据挖掘技术就被越来越多的 CRM 厂商所采用, 例如国外的 Siebel、SAS、PeopleSoft。相对而言目前国内采用数据挖掘技术的分析型 CRM

还较少。因此结合各行业的实际情况研究数据挖掘技术在 CRM 系统中的应用具有重要价值。

本文阐述数据挖掘技术及其在 CRM 中的应用,并通过一个实例说明如何使用数据挖掘技术从大量交易数据和客户信息中提取有价值的规律和模式,帮助企业深入理解客户和改善客户关系。

1 数据挖掘的概念和主要分析方法

1.1 数据挖掘的概念

数据挖掘就是从大量的数据中,抽取出生效的、新颖的、潜在有用的,以及最终可理解的知识、模型或规则的过程^[1]。它是近年来发展起来的一门涉及面很广的交叉学科,融合了人工智能、数理统计、数据库等方面的最新进展,它的应用遍及金融、零售、保险、电信等众多领域,发展前景十分广阔。

1.2 数据挖掘的分析方法

数据挖掘分析方法大致包括关联规则、分类、聚类、序列模式等^[1]。关联规则的目的是希望在事务数据库中发现经常共同出现的项目,推断隐藏在项目之间的相互关联的规则。分类是按照一组数据对象的特征给出其划分的过程。要求有已知分类的样本数据作为训练集,经过对训练集的学习得到关于分类的规律,从而对新数据进行分类。聚类是聚类技术识别一组数据对象的内在规则,将对象聚合成相似的类,以导出数据的分布规律。聚类与分类的差别在于它不需要训练样本,是一种无监督的学习。序列模式的目的是为了挖掘数据之间的联系,但其主要关注于分析挖掘和时间顺序相关的模式。

以上各种分析方法都有各自面向的客观问题,也都有多种实现算法。设计系统的时候,需要根据具体需求选择使用恰当的分析方法以及实现算法。

2 基于数据挖掘技术的分析型 CRM

2.1 CRM 产品的分类

目前业界厂商大多把其 CRM 产品按功能分为操作型、协作型和分析型三类。操作型 CRM 注重通过技术手段实现企业内部与客户相关的销售、服务、支持和市场等业务流程的自动化;协作型 CRM 目的是通过自动化手段集成管理企业所有与客户交互的渠道,改进客户服务;分析型 CRM 目的是在数据挖掘和数据仓库技术的支持下,通过对海量数据的分析,理解客户的分类、价值、忠诚度、消费行为等,帮助企业改善客户关系。分析型 CRM 满足了企业期望深入了解客户的本质需求,无疑将在 CRM 市场上占据日趋重要的地位。

2.2 数据挖掘技术在分析型 CRM 中的主要应用

就当前客户关系管理的应用实践来看,数据挖掘技术在客户关系管理系统中较成熟的应用包括以下几个方面^[2]:(1) 客户群体细分:把客户根据其性别、收入、交易行为特征等属性细分为具有不同需求和交易习惯的群体,同一群体中的客户对产品的需求以及交易心理等方面具有相似性,而不同群体间差异较大。那么这就有助于企业在营销中更加贴近顾客需求。分类和聚类等挖掘方法可以把大量的客户分成不同的类(群体),适合于用来进行客户细分。(2) 交叉营销。交叉营销是指向已购买商品的客户推荐其它产品和服务。这种策略成功的关键是要确保推销的产品是用户所感兴趣的,有几种挖掘方法都可以应用于此问题,关联规则分析能够发现顾客倾向于关联购买哪些商品。聚类分析能够发现对特定产品感兴趣的客户群。

神经网络、回归等方法能够预测顾客购买该新产品的可能性。(3) 客户流失分析。分类等技术能够判断具备哪些特性的客户群体最容易流失，建立客户流失预测模型。从而帮助企业对有流失风险的顾客提前采取相应营销措施。

其它的重要应用还包括客户价值度分析、客户获取、欺诈行为发现等。此处不作详述。

2.3 基于数据挖掘的分析型 CRM 系统的框架结构

我们设计的基于数据挖掘的分析型 CRM 系统包括如下模块：总控程序、图形用户界面、任务组件模块、挖掘引擎及数据挖掘算法函数库、数据准备模块等。图 1 描述了该系统的框架结构和各模块的任务。其中，总控程序负责根据与用户交互的结果选择并控制相应任务组件来完成任务。此外，它也负责控制任务组件管理器和挖掘函数库管理器对库中的任务组件和挖掘函数进行装入、卸载、维护等操作。任务组件根据不同的需求控制数据准备模块进行数据提取、清理等准备工作，控制挖掘引擎从函数库中调用各种数据挖掘算法完成用户指定的客户细分/客户流失分析/交叉销售等分析任务，并把所得模型通过图形界面模块显示给用户从而进行评估和修改。

数据挖掘算法库集成了多种挖掘算法，可以对数据进行关联规则、聚类、分类、预测和偏差检测等分析。数据准备模块通常有如下任务：消除噪声数据或存在不一致的数据，并把来自不同数据源的数据集成在一起，变换为适合挖掘的统一形式，并按要求从中检出与任务相关的数据。数据准备是否充分，对于挖掘算法的效率乃至正确性都有关键性的影响。图形用户接口模块是系统与用户进行交互通讯的模块。用户通过该模块限定查询任务并给出具体需求。该模块也应以可视化的形式给出挖掘的结果，并支持用户对挖掘出的模式进行评估。

整个系统以数据挖掘算法为核心，对企业数据仓库中的数据进行集成与分析挖掘，发现与客户相关的规律与模式。

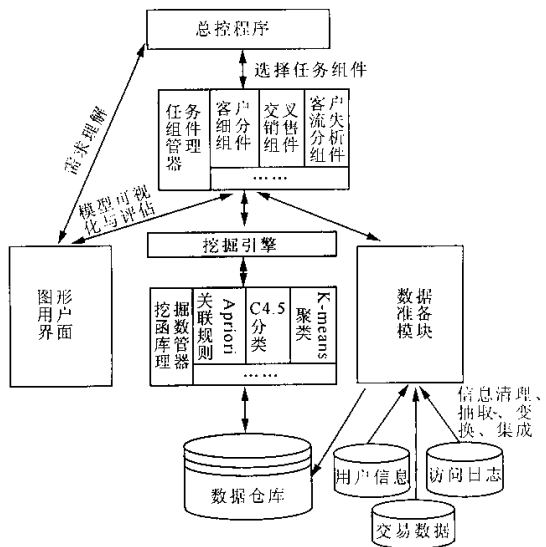


图 1 系统的框架结构

3 基于决策树的客户流失分析

下面在某电子商务系统的客户关系管理子系统中，使用数据挖掘技术来进行客户流失分析，并按照解决数据挖掘问题的一般过程来分步叙述客户关系管理系统的数据挖掘工具的设计与实现。

3.1 需求理解与模型选择

客户流失分析就是该企业希望能够基于数据挖掘技术建立客户流失预测模型，了解哪些因素对客户的流失率有较大影响，从而采取相应措施。分类和聚类都可以用来进行客户流失

分析。分类和聚类的区别在于分类需要已标定的样本,聚类不需要。鉴于可以得到以往客户是否流失的数据用作训练样本,这里我们选择分类方法。

3.2 数据准备

在对数据进行挖掘之前,必须经过数据准备的阶段。数据准备是否充分,对于挖掘算法的效率乃至正确性都有关键性的影响。数据准备阶段用的时间占整个挖掘过程总时间的60%。数据准备通常包括数据清理、集成、变换、归约等工作。

企业的客户信息和交易信息分别存放在用户索引表、个人用户表以及订单表等多个表内。数据准备时必须去除这些表中不需要的信息,例如用户电话、传真、身份证号码、联系方式等,将需要的信息抽取整理成为能被挖掘算法所利用的数据表格(表1)放入挖掘用数据仓库。这也有利于提高算法执行效率。

表1 经抽取而成的客户信息

ID	性别	年龄(岁)	地区	学历	月收入(元)	日均交易频率	交易总金额(元)	...	是否流失
1	男	27	上海	高中	1200	2	97	...	Y
2	女	22	北京	本科	2500	12	651	...	N
...

抽取得到的表中数据的类型和挖掘算法需要的类型不一定一致,还需要做一些转换工作。例如表1的数据需要作一种离散化变换成表2的客户信息。离散化可以由程序自动寻找离散化方法,但是数据自动分段的边界常常显得不够自然,不符合人为习惯,最好能由分析人员根据经验视具体情况来对属性值进行分段,实现离散化。

表2 经离散化变换后的客户信息

ID	性别	年龄(岁)	地区	学历	月收入(元)	日均交易频率	交易总金额(元)	...	是否流失
1	男	26~30	上海	高中	1000~1500	0~5	0~100	...	Y
2	女	21~25	北京	本科	2000~3000	10~20	500~1000	...	N
...

此外,所得到的用户数据很难做到完整全面,用户可能选择不填调查表中的几项甚至全部,造成数据项空缺。对于空缺的数据项,要视情况排除,或填入缺省值,例如对于数值型数据来说可以取平均值作为缺省值。

3.3 核心算法

数据准备完成之后,数据挖掘引擎就可以调用相应的算法来对数据进行挖掘分析。这里我们采用分类的决策树方法来进行客户流失分析。决策树分类有CLS、ID3、C4.5、CART等许多算法。这里我们采用C4.5算法^[3]。

对已经数据准备模块处理好的数据进行测试,就可得到根据该客户的属性和交易行为历史判断该客户是否将会流失的决策树。

3.4 模型评估

决策树算法的执行结果是一棵表明哪些因素影响客户流失的决策树。系统以图形可视化的方式并辅以文本、报表等给出结果,以便于分析人员评估。根据决策树各个分叉节点的测试条件,我们就可以得到有价值的规律。为了得到最有价值的结果,往往可能需要反复多次

地进行数据挖掘以改进模型,直到找出分析人员认可的有价值的结果为止。

得到最终的决策树模型之后,我们就可以使用该模型来对企业的所有客户进行分析预测,找出可能流失的客户群体的组成和特征,并根据预测结果对可能流失的客户采取恰当的营销措施以增强其忠诚度。

4 结束语

客户资源是企业的生命,保留并巩固企业的客户资源对企业来说意义重大。将数据挖掘技术应用于客户关系管理,能够帮助企业深入理解客户,得到更加准确的客户模型,从而改进营销决策和客户服务,具有十分重要的应用价值。随着数据挖掘技术的不断成熟,基于数据挖掘的分析型客户关系管理系统无疑也将获得越来越广泛的应用。

参考文献

- 1 Jiawei Han, Micheline Kamber. 数据挖掘概念与技术. 范明, 孟小峰译. 北京:机械工业出版社, 2001.
- 2 Alex Berso, Stephen Smith, Kurt Thearling. 构建面向CRM的数据挖掘应用. 贺奇, 郑岩, 魏黎等译. 北京:人民邮电出版社, 2001.
- 3 Quinlan, Ross J. C4. 5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

(责任编辑:邓大玉 曾蔚茹)

(上接第254页)

但这种算法在高密度数据集上会存在过适应性,此时效果反而会差一些。

4 结束语

基于项目评分预测的主动营销算法根据项目之间的相似性初步预测用户对未评分项目的评分,在此基础上采用一种新颖的相似性度量方法计算目标用户的最近邻居。实验结果表明,本算法可以有效解决用户评分数据极端稀疏情况下传统相似性度量方法存在的问题,显著提高主动营销系统的推荐质量。模型目前还没有考虑信息量少的用户的推荐,没有考虑模式的动态更新,没有考虑专家对推荐系统的指导作用,这些问题将有待于进一步讨论。

参考文献

- 1 Yoon Ho Cho, Jae Kyeong Kim, Soung Hie Kim. A personalized recommender system based on Web usage mining and decision tree induction. Expert Systems with Applications, 2002, 23: 329~342.
- 2 Gerard Rodriguca Mula, Hector Garcia Molina, Andreas Paepacke. Collaborative value filtering on the Web. Computer Network and ISDN Systems, 1998, 30: 736~738.
- 3 Dong seop Lee, Gye Young Kim, Hyung Il Choi. A Web-based collaborative filtering system. Pattern Recognition, 2003, 36: 519~526.
- 4 John S Breese, David Heckerman, Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998.

(责任编辑:邓大玉 曾蔚茹)