第29卷 第1期 昆明理工大学学报(理工版) Vol.29 No.1
2004年2月 Journal of Kunming University of Science and Technology (Science and Technology) Feb. 2004

CRM数据挖掘中关联规则的应用

李冠乾，许 亮
(昆明理工大学 管理与经济学院，云南 昆明 650093)

摘要：客户关系管理是数据挖掘重要的应用领域，包括客户获取、客户保持、客户价值提升的客户关系管理的各个方面.目前，面向客户关系管理的数据挖掘应用研究涉及应用功能、应用方法、算法、模型、数据处理、系统设计和开发等方面.阐述了数据挖掘在客户关系管理中的应用，通过对关联规则的具体分析，进一步明确数据挖掘在企业经营管理中的重要性.

关键词：数据挖掘；客户关系管理；关联规则
中图分类号：TP392 **文献标识码**：A **文章编号**：1007－855X(2004)01－0113－05

Application of Association Rule in the Data Mining of CRM

LI Guan-qian, XU Liang
(Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Customer Relation Management (CRM) is an essential field of implementation of Data Mining (DM), which includes all aspects of CRM such as the Customer Acquirement, the Customer Maintenance and the Increase of The Value of Customer etc. Nowadays, the research of DM in CRM is involved in the function, method, algorithm, modeling, data processing, system design and development etc. The focus is on the application of DM in CRM. Through the specific analysis of Association Rule, the important role of DM in the management of enterprises is further emphasized.

Key words: data mining; customer relationship management; association rule

0 引 言

数据挖掘已经被Gartner Group列为“未来3～5年内将对工业产生深远影响的五大关键技术”之首，而CRM是它的重要应用领域，因为有了数据挖掘技术的支持，才使CRM的理念和目标得以实现，满足现代电子商务时代的需求和挑战.

1 CRM简介

1.1 CRM的历史

CRM(Customer Relationship Management)客户关系管理，是伴随着因特网和电子商务的大潮进入中国的.最早发展客户关系管理的国家是美国，Gartnet Group首先提出了CRM的概念，认为所谓的客户关系管理就是为企业提供全方位的管理视角；赋予企业更完善的客户交流能力，最大化客户的收益率.

在1980年初便有所谓的“接触管理”(Contact Management)，专门收集客户与公司联系的所有信息；20世纪90年代初期，客户关系管理体现为销售力量自动化系统(SFA)、客户服务系统(CSS)；1996年发展为集销售、服务于一体化的呼叫中心(call center)；1998年，随着电子商务的兴起，CRM开始向eCRM方向发展.

1.2 CRM在中国的发展

在过去的一年多来，我们看到，客户关系管理(CRM)在中国正以迅猛的速度普及.CRM强调“以客户

收稿日期：2003－05－12.
第一作者简介：李冠乾(1978～)，男，在读硕士；主要研究方向：管理科学与工程.

为中心"的管理方法,将客户,而非产品,放在提高企业竞争力的中心位置,这一思想非常适合正在急于寻找不同于价格战、广告战的竞争策略的中国企业.与其他的管理软件,例如 MRP、ERP 的发展历程相比,CRM 被中国企业接受和应用的速度,以及行业渗透的深度和广度都是前所未见的.

1.3 CRM 的作用

CRM 是借助先进的信息技术和管理思想,通过对企业业务流程的重组来整合客户信息资源,并在企业的内部实现客户信息和资源的共享,为客户提供 one - to - one 个性化服务、改进客户价值、满意度、赢利能力以及客户的忠诚度,保持和吸引更多的客户,最终实现企业利润最大化.另一方面,CRM 应用系统通过对所收集的客户特征信息进行智能化分析,为企业的商业决策提供科学依据.

CRM 既是一套原则制度,也是一套软件和技术.CRM 应用软件将最佳的实践具体化并使用了先进的技术来协助各企业实现缩减销售周期和销售成本、增加收入、寻找扩展业务所需的新的市场和渠道、以及提高客户的价值、满意度、赢利性和忠实度等目标.CRM 在整个客户生命期中都以客户为中心,这意味着 CRM 应用软件将客户当作企业运作的核心.CRM 应用软件简化协调了各类业务功能(如销售、市场营销、服务和支持)的过程并将其注意力集中于满足客户的需要上.CRM 应用还将多种与客户交流的渠道,如面对面、电话接洽以及 Web 访问协调为一体,这样,企业就可以按客户的喜好使用适当的渠道与之进行互动性交流.

因特网和电子商务带动了全球的经济,日益激烈的市场竞争正在驱使着企业公司重新确定自己商业模式,并且进一步加强与巩固客户的关系.根据 IDC 2000 年 8 月份的研究报告,全球的 CRM 市场将从 1999 年的 32 亿美元,以年复合增长率 29%的速度增加到 2004 年的 121 亿美元.META Group 则预计全球 CRM 市场年复合成长率为 50%,从 2000 年 130 亿成长达 2004 年 670 亿美元.可见,自 Gartner Group 提出 CRM 概念以来,CRM 应用在全球市场中呈现出迅猛发展的势头.

1.4 CRM 的组成部分

从体系结构角度看,整个 CRM 架构可以分为三个关键部分:

1) 操作层次的 CRM.用于自动地集成商业过程,包括对销售、营销和客户服务三部分业务流程的信息化,客户接触点、渠道、前后端的集成.

2) 客户互动.关注客户接触点的交互,即与客户进行沟通所需的手段(如呼叫中心、网络、电话、E-MAIL 等)的集成和自动化处理.

3) 分析层次的 CRM.用于操作层次 CRM 和客户互动产生的信息的分析处理,通过基于数据仓库的数据挖掘产生商业智能以支持企业战略战术的决策,包括:客户服务支持、客户市场细分、客户变动分析、交互和垂直销售分析、新客户模型、客户接触最优化、广告分析、信用风险分析、客户生命周期价值模型等.

2 CRM 中数据挖掘的应用

2.1 数据挖掘的定义

数据挖掘(Data Mining)就是从大量的、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.CRM 中的数据挖掘就是利用数据挖掘理论和技术创建描述和预测客户行为的模型,以实现企业有效的客户关系管理.

2.2 CRM 中应用的数据挖掘技术

在 CRM 中数据挖掘技术都有着广泛的应用,主要体现在:(1)概念/类描述.概念描述以简洁汇总的形式描述给定的任务相关数据集,提供数据价值的一般特性,一般应用于 CRM 中的描述式数据挖掘.概念或类描述由特征比和比较或区分组成,有两种一般方法:基于数据立方体 OLAP 的方法和面向属性归纳的方法.(2)关联分析.关联分析发现关联规则,广泛用于购物篮、商务管理和决策分析,是商业分析中应用最为广泛的一种数据挖掘方法和模式.(3)分类和预测分析.分类和预测是 CRM 中数据分析的两种重要形式,可以用于提取描述重要数据类的模型或预测未来的数据趋势.主要方法包括:决策树/判定树、贝叶斯法、BP 神经网络算法、遗传算法、粗糙集、模糊集等.(4)聚类分析,属于无指导学习.对象根据最大化类内的相

似性的原则进行聚类或分组.

3 关联分析在 CRM 中的应用

3.1 关联规则的作用以及重要性

传统的企业信息系统由于缺乏数据挖掘功能,最多只有一些统计数据,从表面上似乎合理,但实际上根本不能反映本质情况,例如,通过传统的信息系统,我们得出某一种乳酪产品在某超市的销售额排名第209位,按照以前的做法,该乳酪必定会停止出售,但是通过对数据进行关联分析,我们发现消费额最高的客户中有25%常常买这种乳酪,如果停止出售这种乳酪,必然会引起这些客户的不满.

关联分析是发现交易数据库中不同商品(项)之间的联系,这些规则找出顾客购买行为模式,如购买了某一商品对购买其他商品的影响,例如,他能发现数据库中形如"90%的顾客在一次购买活动中购买商品A的同时购买商品B"之类的知识.发现这样的规则可以应用于商品货架设计、货存安排以及根据购买模式对用户进行分类.而为了达到这个目的,关联分析需要通过关联规则进行数据挖掘.关联规则是Rakesh Agrawal等人提出的数据挖掘领域中的一个重要课题, Agrawal等于1993年首先提出了挖掘顾客交易数据库中项集间的关联规则问题,以后诸多的研究人员对关联规则的挖掘问题进行了大量的研究.他们的工作包括对原有的算法进行优化,如引入随机采样、并行的思想等,以提高算法挖掘规则的效率;对关联规则的应用进行推广.

用于关联规则发现的主要对象是事务型数据库,其中针对的应用则是售货数据,也称货篮数据.一个事务一般由如下几个部分组成:事务处理时间 ,一组顾客购买的物品,有时也有顾客标识号(如信用卡号).

由于条形码技术的发展,零售部门可以利用前端收款机收集存储大量的售货数据.因此,如果对这些历史事务数据进行分析,则可对顾客的购买行为提供极有价值的信息.例如,可以帮助如何摆放货架上的商品(如把顾客经常同时买的商品放在一起),帮助如何规划市场(怎样相互搭配进货).由此可见,从事务数据中发现关联规则,对于改进零售业等商业活动的决策非常重要.

3.2 关联规则的定义

设 $I = \{i_1, i_2, \cdots, i_m\}$ 是 m 个不同项目的集合,给定一个交易数据库 D,其中每一个交易 T 是 I 中一组项目的集合,即 T 为 I 的子集,一条关联规则就是形如 $X \Rightarrow Y$ 的蕴涵式,其中 X、Y 都为 I 的真子集,而且 $X \cap Y = \phi$.

1) 称物品集 X 具有大小为 s 的支持度,如果 D 中有 $s\%$ 的事务支持物品集 X;

2) 称关联规则 $X \Rightarrow Y$ 在事务数据库 D 中具有大小为 s 的支持度,如果物品集 $X \cup Y$ 的支持度为 s;

3) 称规则 $X \Rightarrow Y$ 在事务数据库 D 中具有大小为 c 的可信度,如果 D 中支持物品集 X 的事务中有 $c\%$ 的事务同时也支持物品集 Y.

如果不考虑关联规则的支持度和可信度,那么在事务数据库中存在无穷多的关联规则.

事实上,人们一般只对满足一定的支持度和可信度的关联规则感兴趣.在文献中,一般称满足一定要求的(如较大的支持度和可信度)的规则为强规则.因此,为了发现出有意义的关联规则,需要给定两个阈值:最小支持度和最小可信度.前者即用户规定的关联规则必须满足的最小支持度,它表示了一组物品集在统计意义上的需满足的最低程度;后者即用户规定的关联规则必须满足的最小可信度,它反应了关联规则的最低可靠度.

在实际情况下,一种更有用的关联规则是泛化关联规则.因为物品概念间存在一种层次关系,如夹克衫、滑雪衫属于外套类,外套、衬衣又属于衣服类.有了层次关系后,可以帮助发现一些更多的有意义的规则.例如,"买外套 $\Rightarrow$ 买鞋子"(此处,外套和鞋子是较高层次上的物品或概念,因而该规则是一种泛化的关联规则).由于商店或超市中有成千上万种物品,平均来讲,每种物品(如滑雪衫)的支持度很低,因此有时难以发现有用规则;但如果考虑到较高层次的物品(如外套),则其支持度就较高,从而可能发现有用的规则.

另外,关联规则发现的思路还可以用于序列模式发现.用户在购买物品时,除了具有上述关联规律,还有时间上或序列上的规律,因为,很多时候顾客会这次买这些东西,下次买同上次有关的一些东西,接着又买有关的某些东西.

3.3 APRIORI 算法

Apriori 算法是 Agrawal 等提出的挖掘关联规则的一个重要方法,它可以分解为两个子问题:

1) 找到所有支持度大于等于最小支持度的项目集(Itemset),这些项目集的集合称为频集(Frequent Itemset),记为 $L(k)$,k 代表每个项目集里所含元素个数.

2) 使用第 1 步找到的频集产生期望的规则.其核心思想如下:

F(1) = {frequent 1 - itemsets} //首先产生频集 F(1),它由候选集 C(1)而来(候选集是一个潜在的频集),C(1)中每个项目集只有一个元素(以商店的数据库为例,每个元素代表一种商品),通过把 C(1)中支持度小于最小支持度的项目集删除就得到 F(1),F(1)中的项目集是按一定的次序来排列的.

```
K = 2
While (F(k - 1)非空)
{
C(k) = Apriori_generate(F(k - 1))
//由频集 F(k - 1)产生候选集 C(k),起产生过程如下:
//p,q 分别记为频集 L(k - 1)p,L(k - 1)q 的某一个项目集
//Insert into C(k)
// (select p.item(1), p.item(2),…,p.item(k - 2), p.item(k - 1), q.item(k - 1)
//from L(k - 1)p, L(k - 1)q
//where p.item(1) = q.item(1),…,p.item(k - 2) = q.item(k - 2),p.item(k - 1) < q.item(k - 1) )
```

//以上语句说明,从 L(k - 1)p 的第一个项目集开始,在 L(k - 1)q 中找出所有这样的项目集,使得 p,q 中的前 k - 2 个元素分别对应相等,而 p 中的第 k - 1 个元素要小于 q 中的第 k - 1 个元素(这里的大小是按元素在 F(1)中相对次序来说的),然后把集合{ p.item(1),p.item(2),…,p.item(k - 2),p.item(k - 1),q.item(k - 1) }作为 C(k)中的一个项目集,一直到 L(k - 1)p 的最后一个项目集.这样得到的 C(k)还要经过剪枝后才能成为真正的 C(k),所谓的剪枝就是找出每个项目集的所有子集(该子集由 k - 1 个元素组成)在这些子集中至少有一个不是频集 L(k - 1)的项目集,然后把这些项目集从 C(k)中删除.例如:

存在 L(2)如下

项目集	支持度/%
AB	60
AC	100
AD	80
AE	40
BC	60
BD	40
CD	80
CE	40
DE	40

由 L(2)可得 C(3)如下

项目集	支持度/%
ABC	?
ABD	?
ABE	?
ACD	?
ACE	?
ADE	?
BCD	?
CDE	?

经过剪枝后 C(3)变为(由于项目集 ABE 中的子集 BE 不是 L(2)的项目集,所以要把 ABE 给删除)

项目集	支持度/%
ABC	?
ABD	?
ACD	?
ACE	?
ADE	?
BCD	?
CDE	?

```
for all transactions t in D
  {
    subset(C(k),t) //计算 C(k)中每个项目集的支持度,它由每个项目集在数据库中出现的次数与总纪录数得商得出
  }
```

F(k) = { c in C(k) | c.count > = 最小支持度}　把 C(k)中那些支持度大于等于最小支持度的项目集作为 F(k)的一个项目集

K + + ;

}

结果 = ⋃(n = 1 to k)F(n)

3.4 APRIORI 算法的具体应用

现以购物篮为例，简单说明关联规则在数据挖掘中的应用：

在某商店里有多种货物，例如，钉子、锤子、钳子等，我们应用关联规则就是为了发现当一个人已经购买了锤子时，那他有多大可能还会买钉子呢？一种简单的方法就是运用条件概率，其定义如下：

设 A、B 是两个事件，且 $P(A) > 0$ 称 $P(B/A) = P(AB)/P(A)$ 为在事件 A 发生的条件事件 B 发生的条件概率.

首先，我们设定最小支持度为 40%，假设该商店数据库中共有 5 条记录，如下表所示：

记录中 1 代表购买了，0 代表没有购买

记录号	锤子	钉子	钳子
1	1	1	0
2	0	1	0
3	1	1	0
4	1	0	1
5	0	1	0

根据 Apriori 算法，我们首先产生第一频集 L_1

项目集 X	支持度/%
锤子	60(3/5)
钉子	80(4/8)
钳子	20(1/5)

由于钳子的支持度小于 40%，最终 L_1 频集为

项目集 X	支持度/%
锤子	60
钉子	80

由 L_1 频集，我们得出 C_2 候选集

项目集 X	支持度/%
锤子，钉子	?

由 C_2 候选集，我们得出 L_2 频集

项目集 X	支持度/%
锤子，钉子	40(2/5)

L_2 为我们得出的最终频集

由条件概率可得：

规则“锤子”⇒“钉子”(买了锤子后又买钉子)的可信度 = P(锤子和钉子)/P(锤子) = 40%/60% = 2/3

规则“钉子”⇒“锤子”(买了钉子后又买锤子)的可信度 = P(锤子和钉子)/P(钉子) = 40%/80% = 1/2

我们可以看到买锤子的人也买钉子的可能性(67%)高于买钉子的人也买锤子的可能性(50%).锤子和钉子关联的支持度已经足够高了，意味着这是一条有意义的关联规则.

4 结束语

数据挖掘技术在以客户为中心的电子商务时代扮演着越来越重要的角色，随着理论的进一步发展和深化，必然会带给 CRM 更为广泛的应用前景和市场价值，提高企业的竞争力.

参考文献：

[1] 张喆，常桂然，黄小原.数据挖掘在 CRM 中的应用[J].中国管理科学，2003，(2):53.

[2] 吕廷杰，尹涛，王琦.客户关系管理与主题分析[M].北京：人民邮电出版社，2002.82 ~ 85.

[3] 胡运发.数据与知识工程导论[M]北京：清华大学出版社，2003.128 ~ 129.

[4] 蔡伟杰，张晓辉，朱建秋，朱扬勇.关联规则挖掘综述[EB/OL]http://www.dmgroup.org.cn/lw1.htm.