ID3分类算法在银行客户流失中的应用研究

李 霞
(广东外语外贸大学 信息学院,广东 广州 510006)

摘 要:决策树已被成功应用到许多分类问题上,其中ID3是决策树学习的典型算法。文中就该算法在银行客户流失中的应用做了实例研究。叙述了ID3分类算法的原理及其实现算法,并分析了银行客户流失的原因和分类,以一个具体案例详细讲解了ID3分类算法在银行客户流失分析的具体应用流程,包括:数据采样、数据分析、建立模型和模型解释。文中实现ID3算法并作用于银行数据得到一个银行客户流失模型,通过提取模型中的规则对银行预测客户流失特征具有一定的辅助作用。

关键词:决策树;ID3分类算法;银行客户流失;预测模型
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)03-0158-03

ID3 Applying to Loss of Bank Clients

LI Xia
(College of Information, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: Decision tree has been successfully applied to many of the classification issue, ID3 is the representative algorithm. Gives an instance study for ID3 completing the classification of banks forecast, presents the principle and implementation of the ID3 algorithm and analyses the reason and classification of the loss of bank clients. Also explains concretely the approach of how to use ID3 to forecast the loss of bank clients in detail, for instance, data sampling, data analyzing, model getting and model explaining. At the end, gets a model from which by picking up rules helps the bank forecast the loss of clients.

Key words: decision tree; algorithm of ID3; losing of bank clients; model of forecasting

0 引 言

数据挖掘技术[1]已经被应用在很多领域,其中在银行的客户获得、交叉销售(Cross-selling)、客户关怀与保持等方面,数据挖掘都发挥着重要作用。如:数据挖掘能够帮助银行确定客户的特点,从而可以为客户提供有针对性的服务;通过数据挖掘,可以发现购买某类金融产品的客户特征,从而可以扩大业务等。

客户流失[2,3]一直是影响银行利润及其业绩的一个问题,如果能够通过数据挖掘工具在客户流失之前就找到客户的特征,从而采取一定措施挽回这些客户,这将给银行带来实在的帮助。应用于银行客户流失的技术主要集中在:逻辑回归技术[1,4]和分类预测技术[1,4]。文中研究并实现决策树分类算法ID3[5~7],并将其作用于银行数据,生成一个应用于银行客户流失的预测模型,该模型对解决银行客户流失问题具有一定的辅助作用。

收稿日期:2008-06-06
基金项目:广东省自然科学研究重点项目(06Z012)
作者简介:李 霞(1976-),女,讲师,硕士,主要研究领域为数据挖掘。

1 银行客户流失原因分析

1.1 客户流失的定义及分类

所谓客户流失是指客户不再重复购买,或终止原先使用的服务。根据客户流失原因可将流失客户分成以下几种类型:

(1)自然流失。这种类型的客户流失不是人为因素造成的,比如客户的搬迁和死亡等。这样的客户流失是不可避免的,应该在弹性流失范围之内。自然流失所占的比例很小,银行可以通过提供网上服务等方式,让客户在任何地方、任何时候都能方便快捷地使用银行的产品和服务,减少自然流失的发生。

(2)竞争流失。由于竞争对手的影响而造成的流失称为竞争流失。竞争突出表现在价格战和服务战上。如:客户找到了收益更高的产品而转移购买;竞争对手服务质量的提高;竞争对手产品技术手段的更新而使客户转向购买技术更先进的替代产品。

(3)过失流失。过失流失是由于银行自身工作中

的过失引起客户不满意而造成的。比如，企业形象不佳、服务态度恶劣，客户对银行的产品和服务质量感到不满，并通过直接或间接的渠道投诉得不到解决，这些都会使客户转而投向竞争对手。过失流失在客户流失总量中所占的比例较高，但同时也是企业可以通过采取一些有效手段来防止的。

1.2 客户流失原因分析

有市场竞争就有市场退出，在银行之间的竞争过程中，原有客户的流失相当正常，关键在于必须找到客户流失的原因，进而制定有效的控制策略。导致客户流失的因素主要有以下几种：

(1) 金融服务品种单一。金融服务产品的相对单一，不能随时根据市场变化和用户需求，推出新的金融服务品种和调整金融发展战略，必然导致客户的流失。因此完善金融服务品种和手段、提供实时创新的金融产品和增加个性化服务品种有利于银行固定一批优质客户，降低银行的客户流失率。

(2) 服务与客户关怀不够。客户的流失或保留取决于对产品或服务的评价，客户的抱怨和询问如果不能得到妥善的处理会造成他们的离去。要建立多种渠道反馈客户对产品和服务的意见，让他们感觉到自己受到了尊重。这样做不仅可以提高客户的满意度和忠诚度，而且还能从客户那里收集到免费的建议，以便不断改善银行的产品和服务。

(3) 银行内部员工的流失。银行内部员工的流失，可能导致和它长期保持联系的重要客户的流失。频繁的员工流动不仅增加了银行员工培训的成本，还会使客户不得不重新认识和熟悉新的接触对象，这可能增加了他们的不适而导致流失的发生。

(4) 不注重企业形象。良好的企业形象会增加客户的信赖感。银行应该在各方面尽量避免产生负面的社会影响，以优质的产品和多元化的服务、良好的企业文化、完善的售后服务机制和积极进取的企业目标来赢得客户的信赖，从而减少流失的发生。

2 ID3分类算法

ID3算法是所有可能的决策树空间中一种自顶向下、贪婪的搜索方法。ID3搜索的假设空间是可能的决策树的集合，搜索目的是构造与训练数据一致的一棵决策树，搜索策略是爬山法，在构造决策树时从简单到复杂，用信息熵作为爬山法的评价函数。ID3算法的核心是在决策树各级节点上选择属性，用信息增益作为属性选择的标准，使得在每个非叶节点进行测试时能获得关于被测数据最大的类别信息，使得该属性将数据集分成子集后，系统的熵值最小。期望该非叶节点到达这个叶节点的平均路径最短，使生成的决策树平均深度较小。

2.1 ID3算法原理

设S是s个样本的集合，假定类别属性具有m个不同值，定义m个不同类$C_i(i=1,\cdots,m)$。设s_i是类C_i中的样本数，对一个给定的样本集，它总的信息熵值为：

$$I(s_1,s_2,\cdots,s_m)=\sum_{i=1}^{m}p_i\log_2(p_i) \tag{1}$$

其中p_i是任意样本属性C_i的概率，并用S_i/S估计。

设属性A具有V个不同值。可用属性A将S划分为v个子集$\{s_1,s_2,\cdots,s_v\}$；其中s_j包含S中这样一些样本，它们在A上具有a_j值。如果选择A作为测试属性，则这些子集就是从代表样本集S的节点生长出来的新的叶节点。设s_{ij}是子集s_j类别为C的样本数，则根据A划分样本的信息熵为：

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

其中，$I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}p_{ij}\log_2(p_{ij})$；$p_{ij}=\frac{s_{ij}}{|s_j|}$是$S_j$中类为$C_j$的样本的概率。最后用属性$A$划分样本集$S$后所得的信息增益值为：

$$\mathrm{Gain}(A)=I(S_1,S_2,\cdots,S_m)-E(A) \tag{3}$$

2.2 ID3算法实验过程

算法：由给定的训练数据产生一棵决策树。

输入：训练样本集，其中记录主要由离散值属性描述；

候选属性的集合

输出：一棵决策树。

方法：

1) 创建节点N；

2) If samples都在同一个类C then

3) 返回N作为叶节点，以类C标记；

4) If attribute_list为空 then

5) 返回N作为叶节点，标记samples中最普通的类；//多数表决

6) 选择attribute_list中具有最高信息增益的属性作为test_attribute；

7) 标记节点N为test_attribute；

8) For each test_attribute中的已知值a_i；//划分samples

9) 由节点N长出一个条件test_attribute= a_i的分枝；

10) 设S_i是samples中test_attribute= a_i的样本集合；//一个划分

11) If S_i 为空 then

12) Else 加上一个有 Generate_decision_tree(S_i, attribute_list - {test_attribute})返回的节点

3 ID3 分类算法在银行客户流失分析的应用流程

通过构造银行数据,以数据挖掘的 ID3 分类算法对数据进行分类预测,找到那些引起客户流失的规则,通过这些规则指导银行对某些客户某些行为进行一定的特殊管理。

3.1 数据采样

从银行的各业务数据库中采集样本数据。数据选择包括目标变量的选择、输入变量的选择和建模数据的选择等多个方面。

(1)目标变量的选择:在客户流失分析系统中,实际面对的流失主要有账户取消发生的流失和账户休眠发生的流失两种形式。对于不同的流失形式,需要选取不同的目标变量。

(2)输入变量的选择:输入变量用于在建模时作为自变量寻找与目标变量之间的关联。在选择输入变量时,通常选择两类数据:静态数据和动态数据。静态数据指的是通常不会经常改变的数据,包括客户的基本信息(如性别,年龄,婚姻状况,职业,居住地区等)。动态数据指的是经常或定期改变的数据,如每月存取记录、消费金额、消费特征等等。

(3)建模数据的选择:由于银行客户的流失主要是自然流失、竞争流失和过失流失三种,自然流失是由于客户的迁徙等原因导致的客户流失,而竞争流失和过失流失是竞争对手的优惠政策和客户对目前的服务不满意而导致的客户流失,显然第二、三种流失的客户才是银行真正关心的,对银行具有挽留价值的客户。

根据以上分析,在选择建模数据时必须选择第二、三种流失的客户数据参与建模,属性分为:"是否定期","存款数","月业务频率","是否投资","是否流失"。

3.2 数据分析

数据分析就是对采样后的数据进行初步分析,试图寻找出不同变量之间的关联度,以及不同变量对于客户流失的影响程度。并非所有输入变量都是同样的重要,部分因子可能同客户流失无关,删除那些和客户流失概率相关性不大的变量,减少建模变量的数量。这样不仅可以缩短建立模型的时间,减小模型的复杂程度,而且可使建立的模型更加精确。

文中笔者根据人工分析,对于那些与客户流失关系不明显的属性没有给予选择,如姓名、性别、身份证号等等信息。

3.3 建立预测模型

算法作用的部分数据见表 1(未全部列出)。

表 1 部分数据

是否定期	存款数	月业务频率	是否投资	是否流失
"否"	"10000~20000"	"5~10"	"不是"	"不流失"
"否"	"5000~10000"	">10"	"是"	"不流失"
"否"	"20000~30000"	"<2"	"不是"	"流失"
"是"	"10000~20000"	"<2"	"不是"	"不流失"
"否"	"<5000"	"<2"	"不是"	"流失"
"否"	"<5000"	">10"	"是"	"不流失"
"否"	">30000"	">10"	"是"	"不流失"
"是"	"5000~10000"	"<2"	"不是"	"不流失"

针对以上数据,通过 ID3 决策树算法分类后,得到一棵决策树(见图 1)。

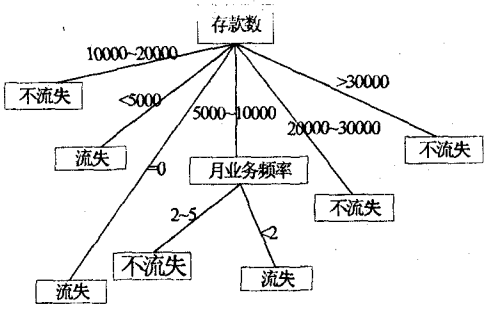

图 1 ID3 算法得到的一棵决策树

3.4 模型的解释与应用

通过分析这棵决策树,可以直接提取出分类规则,当然也可以通过程序直接得到,通过这 20 条数据训练得到有关银行客户流失的规则如下:

1)如果存款数为 10000~20000 之间,则结果为客户不流失;

2)如果存款数为小于 5000 元,则结果为客户流失;

3)如果存款数为 5000~10000 之间,并且如果业务频率为 25 次,则结果为客户不流失;

4)如果存款数为 5000~10000 之间,并且如果业务频率为少于 2 次,则结果为客户流失;

5)如果存款数为 0,则结果为客户流失;

6)如果存款数为 20000~30000 之间,则结果为客户不流失;

7)如果存款数为大于 30000 元,则结果为客户不流失。

通过决策树发现,该银行中的客户流失与客户是否是定期客户或者该客户是否是作为投资没有关系,而与该客户的存款数及月业务频率有着直接的关系。而存款数与月业务频率两者之间,存款数程度更为重要。有了这个预测模型,银行就可以根据该预测模型,

(下转第 167 页)

若想让系统每次启动时都能自动地加载静态ARP,则可将这个批处理软件拖到“windows——开始——程序——启动”中。使用静态ARP缓存增大了网络维护量,在较大或经常移动主机的网络中这样做更为困难。使用静态ARP缓存只能防止ARP欺骗,对IP地址冲突、Flood攻击仍然没有办法阻止。

2)对病毒源头的机器进行处理,杀毒或重装系统。此操作非常重要,因为解决了ARP欺骗的源头PC机的问题,就可以保证内网免受攻击。

3)用可防ARP攻击的交换机(彻底防治)[6]。使用三层交换机,绑定“端口-MAC-IP”,限制ARP流量,及时发现并自动阻断ARP攻击端口,合理划分VLAN,彻底阻止盗用IP、MAC地址,杜绝ARP的攻击,这也是目前防止ARP攻击的最有效方法之一。

5　结束语

网络欺骗攻击作为一种非常专业化的攻击手段,给网络安全管理者带来了严峻的考验。ARP欺骗是一种典型的欺骗攻击类型,它利用了ARP协议存在的安全漏洞,并使用一些专门的攻击工具,使得这种攻击变得普及并有较高的成功率。文中通过分析ARP协议的工作原理,探讨了ARP协议从IP地址到MAC地址解析过程中的安全性,给出了网段内ARP欺骗的实现过程,提出了一种有效的检测系统和几种可行的解决方案,以最大限度地杜绝ARP欺骗攻击的出现。总之,对于ARP欺骗的网络攻击,不仅需要用户自身做好防范工作之外,更需要网络管理员应该时刻保持高度警惕,并不断跟踪防范欺骗类攻击的最新技术,做到防范于未然。

参考文献:

[1] Plummer D C. An Ethernet Address Resolution Protocol[M]. 北京:机械工业出版社,1982.

[2] 谢希仁. 计算机网络[M]. 第4版. 北京:电子工业出版社,2003.

[3] Nachreiner C. Anatomy of an ARP Poisoning Attack[EB/OL]. 1999-07-11. http://www.watchgu-ard.com.

[4] Liruixue. ARP协议的缺陷及ARP欺骗的防范[M]. 北京:机械工业出版社,2007.

[5] 任　侠,吕述望. ARP协议欺骗原理分析与抵御方法[J]. 计算机工程,2003,29(9):127-128.

[6] 宋　志. 一种基于PVLAN的反ARP欺骗的技术实现方法[J]. 计算机安全,2007(10):55-59.

(上接第160页)

跟踪和发现客户的流失趋势,及早采取预防措施,最大限度地降低客户流失率。

4　结束语

研究和实现了决策树分类算法ID3,通过该算法作用于银行数据,得出一个银行客户流失的模型,通过提取模型中的规则,对于银行预测客户流失特征具有一定的辅助作用。

参考文献:

[1] Han Jiawei, Kamber M. Data mining: Concepts and Technique [M]. Beijing: China Machine Press, 2006.

[2] 盛昭瀚,柳炳祥. 客户流失危机分析的决策树方法[J]. 管理科学学报,2005,8(2):20-25.

[3] Tan Pang-Ning, Steinbach M, Kumar V. 数据挖掘导论[M]. 北京:人民邮电出版社,2006:89-193.

[4] Rud O P. 数据挖掘实践[M]. 北京:机械工业出版社,2003:225-264

[5] 王黎明. 决策树学习及其剪枝算法研究[D]. 武汉:武汉理工大学,2007.

[6] 任　伟,丁荣涛. 改进的ID3算法在学习模型的研究与应用[J]. 福建电脑,2007(8):109-110.

[7] 杨　明,张载鸿. 决策树学习算法ID3的研究[J]. 微机发展(现更名:计算机技术与发展),2002,12(5):6-8.

(上接第163页)

海交通大学,2007.

[4] 任　栋,刘连忠. 一种Web应用环境下安全单点登录模型的设计[J]. 计算机工程与应用,2002,38(24):174-176.

[5] 黄　建,倪惜珍. 引入时间特性的角色访问控制[D]. 北京:中国科学院研究生院,2003.

[6] Matheus A. How to declare access control policies for XML structured information objects using OASIS' eXtensible Access Control Markup Language (XACML) [C]//. proceedings of the 38th Hwaii Conference on System Sciences. [s. l.]: [s. n.], 2005.

[7] Chou Shih-Chien. LRBAC: A Multiple-Levelled Role Based Access Control Model for Protecting Privacy in Object-Oriented Systems[J]. Journal of Object Technology, 2004, 3(3): 91-120.

[8] 许　谦,雷咏梅. 一种增强访问控制的服务发现机制[J]. 计算机技术与发展,2007,17(5):99-100.