客户流失危机分析的决策树方法①

盛昭瀚[1]，柳炳祥[2]
(1. 南京大学工程管理学院，南京 210093；2. 东南大学经济管理学院，南京 210018)

摘要：叙述了客户流失的基本概念，分析客户流失危机产生的原因，指出了传统的客户流失危机分析方法存在的问题，在 ID3 算法的基础上，提出一种加权熵的概念，并对 ID3 算法进行改进，提出了一种基于决策树的客户流失危机分析方法，通过一个客户流失危机的分析实验对该模型进行了验证. 实验结果表明该方法是可行的和有效的，为客户流失危机分析和预警提供了一种新的研究思路和分析方法.

关键词：客户流失危机；决策树；ID3 算法；加权熵
中图分类号：TP311　　**文献标识码**：A　　**文章编号**：1007 - 9807(2005)02 - 0020 - 06

0 引　言

危机管理及其预警是管理科学与工程领域中的一个前沿和热点问题，是多学科交叉的边缘性研究课题. 由于危机产生的原因复杂，种类繁多，是带有大量不确定因素的半结构化问题或非结构化问题，很难科学地计算和评估. 传统的分析方法在解决企业危机预警方面存在很大的局限性，分析和预测的结果也不够理想，迫切需要运用其它技术和分析方法来研究危机的分析和预警问题. 国内外已有学者运用数据挖掘技术进行危机管理及其预警方面的研究[1~7]，主要集中在金融危机、财务危机分析和预警上，但对经营危机分析和预警研究还不够深入. 本文利用数据挖掘中的决策树方法研究企业的客户流失危机.

1 客户流失危机分析

1.1 客户流失概念

客户流失是指客户终止与本公司的服务合同或转向其它公司提供的服务[8]. 由于各种因素的不确定性和市场不断的增长以及一些竞争对手的存在，很多客户不断地从一个供应商转向另一个供应商只是为了求得更低的费用以及得到更好的服务，这种客户流失在许多企业中是普遍存在的问题. 因客户流失导致的损失是巨大的，因为获取一个新客户，要在销售、市场、广告和人员工资上花费很多的费用，而且大多数新客户产生的利润不如那些流失的客户多，因此，保留住客户，防止因客户流失而引发的经营危机，对于提高公司的竞争力具有战略意义.

1.2 客户流失危机原因

产生客户流失危机的根本原因是企业客户关系管理过程中存在的各种不确定性，即各种风险. 所谓客户流失风险是在一个给定的时期内，随着企业客户关系管理中的各种参数和基本变量的变化，企业客户流失可能产生的概率分布，离散程度越大，企业面临的客户流失风险就越大. 根据客户流失危机产生的根源不同，可以将危机产生的原因划分为内部因素和外部因素[8]. 内部因素是在一定外部经营环境之下，由于创新能力下降、经营不善、观念滞后、战略决策失误等原因导致企业客户关系管理工作的失败；外部因素是企业的外部经营条件，如政治、经济政策、科技发展、市场需求

① 收稿日期：2001 - 12 - 14；修订日期：2004 - 12 - 04.
基金项目：国家自然科学基金资助项目(79830010).
作者简介：盛昭瀚(1944 —)，男，江苏镇江人，教授，博士生导师.

和竞争条件的突变或恶化,危及企业原有的平衡,使客户关系管理工作出现严重困难. 内部因素可控性较强,是企业自身矛盾的结果;外部因素几乎难以控制,是经营环境矛盾的结果,但客户流失危机往往又是二者综合的结果和体现.

1.3　传统分析方法存在的问题

客户流失危机的分析模型有统计分析法、主观概率法、指标分析法和指数法[9]. 统计分析法是对客户流失危机发生的历史数据进行统计,用定量预测的方法对这些数据进行处理,获取危机发生的概率的方法;主观概率法是对事物发展的概率分布所作的主观判断,是一种运用专家的经验,对客户流失危机的发生概率进行主观判断,并以此为基础做出概率预测的方法;指标分析法是采用各种方法分析和评价各种定性和定量的客户流失危机指标,为企业提供早期的客户流失预警信息;指数法是利用危机指数来反映危机对人们的现实威胁的综合性指标,是危机危害度和发生概率两个因素的乘积. 由于客户流失危机产生的原因复杂,需要对企业的外部环境、内部经营状况进行分析,一个或多个外部环境因素的变化,一个或多个商务进程或系统的不规范或不完美,商务进程和营销活动管理不良,信息处理过程中的错误,管理行为不当引发的事件等,很难科学地计算和评估,只能采用定性和定量相结合的方法确定. 借助上述分析模型很难对客户流失危机进行有效的评价,本文采用数据挖掘中的决策树方法来分析客户流失危机.

2　决策树算法

2.1　决策树方法

决策树方法是利用信息论中的信息增益寻找示例数据库中具有最大信息量的属性字段,建立决策树的一个节点,再根据该属性字段的不同取值建立树的分枝,然后在每个分枝重复递归建立树的下一个节点和分枝的过程[10]. 因为决策树方法是利用信息论原理对大量实例的特征进行信息量分析,计算各特征的信息熵,找出反映类别的重要特征,因此抓住了问题的本质,具有建立的决策树少、分类准确率高、生成的规则简单等特点,应用十分广泛. 例如某公司的客户数据库保存了客户的有关记录,根据其赢利能力大小,将客户分为 3 类,即赢利能力大、一般、小. 示例数据如表 1 所示,对应的决策树如图 1 所示.

表 1　示例数据

Table 1　Example data

客户编号	月收入	年龄	赢利能力
10011	2 000	31	大
10012	2 500	46	一般
10013	1 200	27	小
10014	1 900	25	大
10015	1 650	41	一般

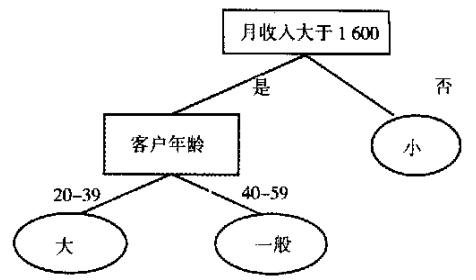


图 1　决策树

Fig. 1　Decision tree

利用决策树进行分析,可以发现一些具有重要商业价值的规则信息,如从上面的例子中可以得出这样的规则:赢利能力大的客户一般月收入大于 1 600 元,年龄在 20～39 岁之间. 因此在企业的营销促销过程中可以有目的、有针对性地做好这类客户群体的服务工作,以极大地提高营销促销工作的效率和企业的盈利水平.

2.2　ID3 算法

国际上最早、最有影响的决策树方法是 Quinlan提出的 ID3 算法[11]. 该算法引入信息论中的信息增益,作为对实体选择重要特征的度量,以信息增益最大的特征产生决策树的结点,由该结点的不同取值建立树的分枝,然后对各分枝递归. 使用该方法建立决策树的结点和分枝,一直到某一子集中的例子属于同一类.

设 $E = F_1 \times F_2 \times \cdots \times F_n$ 是 n 维有穷向量空间,其中 F_j 是有穷离散符号集, E 中的元素 $e = \{V_1, V_2, \cdots, V_n\}$ 称为例子,其中 $V_j \in F_j, j = 1,2,\cdots,n$.

定义 1　正例集和反例集. 设某属性的值为 θ,根据该属性值 θ 将例子 E 分为 2 个子集,其中, $PE = \{V_j \geq \theta, V_j \in F_j\}$, $NE = \{V_j < \theta, V_j \in F_j$,称 PE、NE 为例子 E 的正例集和反例集. 假设向量空间 E 的正例集 PE 和反例集 NE 的大小分别

为 p 和 n,一棵决策树对一例子作出正确类别判断的信息量为

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad (1)$$

定义 2 期望熵:如果以属性 A 作为决策树的根,属性 A 具有 v 个值$\{V_1, V_2, \cdots, V_v\}$,它将 E 分成 v 个子集$\{E_1, E_2, \cdots, E_v\}$,假设 E_i 中含有 P_i 个正例和 N_i 个反例,子集 E_i 所需的期望信息是 $I(P_i, N_i)$,定义以属性 A 为根所需的期望熵为

$$E(A) = \sum_{i=1}^{v}\frac{P_i + N_i}{P + N}I(P_i, N_i) \quad (2)$$

定义 3 信息增益:以属性 A 为根的信息增益为

$$\mathrm{Gain}(A) = I(p,n) - E(A) \quad (3)$$

ID3 算法计算每个属性的信息增益,选择具有最高信息增益的属性 A^* 作为根结点,对 A^* 的不同取值对应的 E 的 v 个子集 E_i 递归调用上述过程生成 A^* 的子结点 $B_1, B_2, \cdots, B_v$.

ID3 算法的基本原理是基于两类分类问题,可以非常容易地将该算法扩展到多个类别. 假使样本集 S 共有 C 类样本,每类样本数为 $p_i(i = 1, 2, \cdots, c)$,若以属性 A 作为决策树的根,A 具有 v 个值,$v = \{V_1, V_2, \cdots, V_n\}$,它将 E 分成 v 个子集 $\{E_1, E_2, \cdots, E_v\}$,$E_i$ 中含有 j 类样本的个数为 P_{ij},$j = 1, 2, \cdots, c$,那么子集 E_i 的信息熵是

$$I(E_i) = -\sum_{j=1}^{c}\frac{p_i}{|E_i|}\cdot\log_2\frac{P_{ij}}{|E_i|} \quad (4)$$

以 A 为根分类的信息熵为

$$E(A) = \sum_{i=1}^{v}\frac{|E_i|}{|E|}\cdot I(E_i) \quad (5)$$

选择属性 A^* 使 $E(A)$ 最小,信息增益 $\mathrm{Gain}(A)$ 将最大. 由此可见,ID3 算法是一个典型的决策树学习系统,它是以信息熵作为分离目标评价函数,采用自顶向下不可返回的策略,搜出全部空间的一部分,以确保决策树建立最简单,构造的决策树平均深度小,分类速度较快. 但 ID3 算法偏向于选取属性较多的属性,而属性较多的属性不一定是最优的属性. 此外,ID3 算法学习简单的逻辑表达能力较差. 针对这些不足,本文提出加权熵的思想,对传统的 ID3 算法进行改进.

2.3 ID3 算法的改进

传统的 ID3 算法选择属性 A 作为新的属性的原则是使 $E(A)$ 最大,并以此为依据产生分枝,这对左右分枝的记录数相差不大的情况下是非常有效的[12]. 如果左右分枝的记录数相差太远,用信息增益来判断可能得不到好的决策树,为了解决这个问题,提出一种称为加权熵的概念.

定义 4 加权熵:设 A 为选择属性,A 有 v 个属性值,对应的权数为 $\omega_1, \omega_2, \cdots, \omega_v$,按照 ID3 算法对属性 A 进行扩展,对应的信息熵为 $E(B_1)$, $E(B_2)$, $\cdots$, $E(B_v)$,定义加权熵为

$$E(A)^* = \sum_{i=1}^{v}\omega_i \cdot E(B_i) \quad (6)$$

式中:$(B_1, B_2, \cdots, B_v)$ 为 v 个结点选择的属性,ω_i 是指分支子集所占的权数. 本文用分枝子集 B_i 在整个集合中所占的比重来计算权数 ω_i,然后计算出加权熵,通过比较加权熵的大小来选择属性的取值. 改进的 ID3 算法的基本步骤如下:

1) 对欲选择的属性 A,假设 A 有 v 个属性值,对应的权数分别为 $\omega_1, \omega_2, \cdots, \omega_v$,以属性 A 为扩展,生成 v 个子结点 $(B_1, B_2, \cdots, B_v)$,按照公式 (2) 求对应的信息熵 $E(B_1)$, $E(B_2)$, $\cdots$, $E(B_v)$;

2) 根据公式(6) 计算加权熵 $E(A)^*$;

3) 选择属性 A^* 使得选择 $E(A^*)^*$,将 A^* 作为新选择的属性;

4) 利用步骤 1) 的计算结果,建立结点 A^* 的后继结点 $(B_1, B_2, \cdots, B_v)$;

5) 对所有的 B_i,若 B_i 为叶子结点,则停止扩展此结点,否则递归执行步骤 1) 至步骤 5),最终完成决策树的建立. 判断决策树构建过程结束的条件有两种情况:给定节点的所有样本都属于同一类;没有多余的属性用来进一步划分样本.

改进的 ID3 算法描述如下[12]:

```
Procedure Algorithm_decision_tree// 根据给定的训练数据产生决策树
Input: 训练样本 example, 候选属性的集合 attribute_list
Output: 一棵决策树
Begin
S = example// 创建节点 N
  if S 都在同一个类 C then
     返回 N 作为叶节点,以类 C 标记
  if attribute_list = NULL then
     返回 N 作为叶节点
  for 样本集 S do
```

for 类 C 中的每个属性 do
利用公式(3) 计算 attribute_list 中具有最大信息增益的属性 max_list
T = max_list// 具有最大信息增益的样本集合
利用公式(6) 计算加权熵 $E(A)^{*}$
endfor
endfor
for each max_list 中的已知值 a_i do
由节点 N 生长出一个分枝
endfor
if T = Null then
加上一个树叶
else 递归执行算法 Procedure Algorithm_decision_tree
End

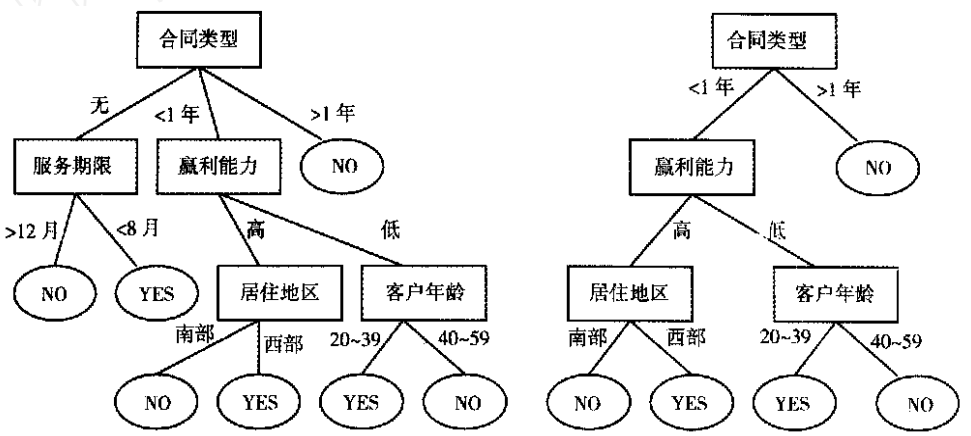
3　在客户流失危机分析中的应用

3.1　实验过程

决策树方法具有效率高、结构简单、易于理解和较高的分类精度的特点,对决策树的 ID3 算法进行改进,引进了加权熵的概念,探讨决策树分类方法在客户流失危机分析和预测中的可行性和有效性,提出了一种基于决策树的客户流失危机分析的方法.

在 Pentium Ⅲ- 733M 联想计算机上用 Visual C + + 实现了 ID3 的改进算法,利用 http://www.ics.uci.edu/mlearn/MlSummary.html 上提供的公用模拟客户数据进行实验,这些数据包括客户的地理分布信息、基本信息、服务信息等[4]. 从某年的客户数据库中通过统计抽样,抽取出 1 000 条记录,以这些数据记录作为训练样本,建立决策树的分类预测模型,以便构造客户流失分析的决策树,分析客户流失主要与哪些因素有关,为客户关系管理中做好客户保留、预防客户流失工作提供科学的参考依据.

在实验中设计了二个方案,一种方案是使用传统的 ID3 算法建立决策树,另一种是使用改进的 ID3 算法建立决策树,建立的客户流失危机分析的决策树如图 2 所示.



(Ⅰ) 方案 1 建立的决策树　　(Ⅱ) 方案 2 建立的决策树

图 2　两种方案建立的客户流失分析决策树

Fig. 2 Decision tree based on custom forfeit analysis

图 2 中,NO 表示客户不流失,YES 表示客户流失. 可以看出,用改进的 ID3 算法构建的决策树比传统的 ID3 算法构建的决策树结构简单,形成的规则容易理解,如 75 % 的客户在第 1 年的合同期满时容易流失,20 % 赢利能力高、年龄段在 20～39 岁的客户发生流失,南部地区的客户流失率低,年龄段在 20～39 之间的客户流失率高,服务期限超过一年的客户 86 % 不会发生流失等规则信息.

3.2　结果分析

使用改进的 ID3 算法,无须对客户数据进行预处理就能对客户数据库进行数据挖掘工作,建立一棵预测客户流失的决策树. 通过该决策树,管

理者能够非常容易地理解和解释从预测模型中发现的一些规则信息,可以深入地了解哪些因素对客户的流失有较大的影响,分析客户为什么会发生流失?什么因素导致客户发生流失?采取何种措施控制客户的流失?两个方案建立的决策树比较表明,利用改进的 ID3 算法建立的决策树,不但可以加快决策树的生长,而且决策树结构简单,便于从中挖掘出易于理解的规则信息.

通过对 ID3 算法的分析知道,该算法时间复杂性为 $O(n)$,空间复杂性为 $O(n)$,采用改进算法虽然计算加权熵需要一定的时间,但随着记录数据的增大,所花费的时间开销会随着决策树的快速生长得到一定程度的弥补,时间复杂性与 ID3 算法基本上处于同一个数量级,因此改进的算法提高了 ID3 算法的效率和性能.

实验结果表明,使用该算法进行客户流失危机的分析和预测是可行的和有效的,它可以帮助管理者更好地了解客户的流失受哪些因素的影响,以便在今后的市场营销中有针对性地对那些流失率高的客户做好服务工作,防止客户的流失引发的经营危机,这对于提高公司的竞争力,改善客户关系而言具有重要意义.

3.3 与其它决策树算法的比较

为了检验改进的 ID3 算法的有效性,用所提供的公用模拟客户数据进行了多项实验,并与一些典型的决策树算法(如 ID3、C4.5、刘小虎提出的 MID3 算法、赵卫东提出的粗集算法[11,7,13,14])进行了分析和比较,决策树不同算法的分类准确率的实验结果如表 2 所示.

表 2 决策树不同算法的分类准确率

Table 2 Rate of sort accuracy on different decision trees (%)

实验数据集	ID3	C4.5	MID3	粗集算法	本文算法
实验数据集 1	77.39	86.16	84.48	83.74	82.63
实验数据集 2	68.32	73.62	72.68	70.55	70.42
实验数据集 3	80.11	87.45	85.36	83.24	84.36
实验数据集 4	75.05	79.62	78.43	81.34	77.42

从表 2 可以看出,改进的 ID3 算法改善了决策树的性能,其分类准确率比传统的 ID3 算法高,与 MID3 算法和粗集算法基本相同,比 C4.5 算法的分类精度稍差.改进的算法克服了 ID3 算法过多地强调选择较多属性的不足,在选取属性时综合考虑了信息增益和分枝子集在整个数据集中所占的比重,虽然该算法的时间复杂性有所增加,但随着数据集的增大,所花费的时间会因为决策树的快速生长得到一定程度的弥补,其时间复杂性与 ID3 算法基本相同[15],它在决策树的效率和精度间作了一定程度的折中.

4 结 论

为了解决决策树中的分枝取值的难点问题,本文在分析传统 ID3 算法的基础上,引入加权熵的概念,对 ID3 算法进行了改进,提出了一种基于决策树的客户流失危机的分析方法,将决策树方法应用于客户流失危机的分析和预测,并利用公用模拟客户数据进行了实验.实验结果表明,该算法是可行的和有效的.利用该算法可以提高决策树的生长速度,优化决策树的结构,挖掘出易于理解的规则信息,为客户流失危机的分析和预防提供了一种新的研究思路和分析方法.

参 考 文 献:

[1]Robert Heath. Looking for answers: Suggestions for improving how we evaluate crisis management[J]. Safety Science, 1998, 30(5): 151—163.

[2] Kuo R J. A decision support system for market through integration of fuzzy neural networks and fuzzy Delphi[J]. Artificial Intelligence, 1998, 12(2): 501—520.

[3]Charles X. Data mining for direct marketing: Problems and solutions[A]. In Proceedings of KDD98: The AAAI-118 Workshop on Knowledge Discovery in Databases[R]. AAAI Technical Report, 1998.

[4]Huges. The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database[M]. Chicago: Irwin Professional, 1996.

[5]赵卫东, 盛昭瀚. 粗集在决策树生成中的应用[J]. 东南大学学报, 2000, 30(4): 132—137.

Zhao Weidong, Sheng Zhaohan. Application of rough sets to the designing of decision trees[J]. Journal of Southeast University, 2000, 30(4): 132—137. (in Chinese)

[6]赵卫东，盛昭瀚. 基于案例推理的决策问题求解研究[J]. 管理科学学报，2000，3(4)：29—36.
Zhao Weidong, Sheng Zhaohan. Decision problem solution based on case-based reasoning[J]. Journal of Management Sciences in China, 2000, 3(4): 29—36. (in Chinese)

[7]赵卫东，李旗号. 粗集在决策树优化中的应用[J]. 系统工程学报，2001，16(4)：289—295.
Zhao Weidong, Li Qihao. Application of rough set to optimization of decision trees[J]. Journal of System Engineers, 2001, 16(4): 289—295. (in Chinese)

[8]James. 企业的泛风险管理[M]. 长春：吉林人民出版社，2001. 46—59.
James. Enterprise Wide Risk Management[M]. Changchun: Jilin People Press, 2001. 46—59. (in Chinese)

[9]苏伟伦. 危机管理[M]. 北京：中国纺织出版社，2001. 57—68.
Su Weilun. Crisis Management[M]. Beijing: The Spinning and Weaving Press, 2001. 57—68. (in Chinese)

[10]CHEN Ming-syan. Data mining: An overview from a database perspective[J]. IEEE Transaction On Knowledge and Data Engineering, 1996, 8(6): 866—883.

[11]Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81—97.

[12]Han Jiawei. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001. 188—195.
Han Jiawei. Data Mining Concept and Technology[M]. Beijing: The Mechanical Engineering Press, 2001. 188—195. (in Chinese)

[13]刘小虎，李　生. 决策树的优化算法[J]. 软件学报，1998，9(10)：797—799.
Liu Xiaohu, Li Sheng. An optimized algorithm of decision tree[J]. Journal of Software, 1998, 9(10): 797—799. (in Chinese)

[14]Quinlan J R. C4.5: Programs for Machine Learning[M]. San Mateo: Morgan Kaufmann Publisher, 1992. 170—227.

[15]Alex Berson. 构建面向 CRM 的数据挖掘应用[M]. 北京：人民邮电出版社，2001. 85—91.
Alex Berson. Building Data Mining Applications for CRM[M]. Beijing: The Post and Telecommunications Press, 2001. 85—91. (in Chinese)

Research method of custom forfeit crisis based on decision tree

SHENG Zhao-han[1], *LIU Bing-xiang*[2]

1. School of Management and Engineering, Nanjing University, Nanjing 210093, China;
2. School of Economic and Management, Southeast University, Nanjing 210018, China

Abstract: This paper tells the basic concept of custom forfeit crisis and analyses the procreant reason resulting from the custom forfeit. It points out the existent problems of the traditional analysis models in evaluating custom forfeit risks. The paper puts forward a research method of custom forfeit crisis based on decision tree. An Algorithm ID3 is described at length, the concept of right entropy is put forward, and the Algorithm ID3 is improved. An analysis experiment of custom forfeit risks on decision tree is illustrated deliberately, experimental results are discussed, and the Algorithm ID3 is compared with other algorithm of decision tree, to prove validity of the Algorithm ID3. The result provides a new research approach for analysis and prevention of custom forfeit risks.

Key words: custom forfeit crisis; decision tree; algorithm ID3; right entropy