分类算法及其在电信客户保持的应用①

左国才, 周荣华, 符开耀

(湖南软件职业学院 软件工程系, 湘潭 411100)

摘 要: 由于电信市场竞争日益激烈, 为了保持客户, 防止客户流失, 提升企业的竞争力, 利用 DBSCAN 算法对流失客户群数据进行划分, 分析客户流失原因, 结合决策树 ID3 算法进行客户流失预测, 实验结果表明, 两种算法的结合, 使得客户流失预测准确率得到较大提高.

关键词: 数据挖掘; DBSCAN 算法; 决策树 ID3 算法; 客户保持

Classification Algorithm and Its Application in Telecom Customer Retention

ZUO Guo-Cai , ZHOU Rong-Hua, FU Kai-Yao

(Software Engineering, Hunan Vocational Institute of Software, Xiangtan 411100, China)

Abstract: Due to an increasingly competitive telecommunications market, in order to keep customers, to prevent the loss of customers, enhance the competitiveness of enterprises, this paper uses DBSCAN algorithm to the loss of customers data types, analysis of customer loss, combined with the decision tree ID3algorithm in customer churn prediction, experimental results show that, two kinds of algorithm combined, makes the customer churn prediction accuracy obtained improved.

Key words: data mining; DBSCAN algorithm; decision tree ID3 algorithm; customer retention

1 引言

竞争格局的改变, 使得电信企业的竞争日益加剧, 企业之间的竞争最终在于客户的竞争. 客户保持的主要目的是防止客户流失. 由于电信市场日趋饱和, 所以获取新客户的成本比留住现有客户要昂贵得多, 竞争对手、技术、策略等动态市场变化, 更容易使客户流失. 因此, 分析客户流失的原因, 找出客户流失的特征, 对存在有流失倾向的客户进行有效预警, 并及时采取有效措施, 尽可能地留住客户, 提高客户保持率成为当前数据挖掘的一个重要研究热点[1].

引起客户流失的原因很多, 单一的客户类别的划分难以准确地建立相应的模型, 使得现有算法在应用分析中的准确度不太理想. 本文采用基于密度的 DBSCAN 算法对流失客户群数据进行划分, 并结合决策树 ID3 算法进行实验, 结果表明两种算法的结合, 使得客户流失预测准确度得到较大提高.

2 电信客户流失分析

在分析客户流失原因时, 应结合客户的价值、消费结构、使用习惯以及消费趋势等, 不同的消费水平流失程度不同, 因此必须对电信客户进行细分, 以客户保持为目的, 从套餐的设计、产品的研发、竞争对手的技术等关键指标, 实现针对不同层次的客户研究相应的营销策略[2].

客户细分首先应该确定客户细分的标准, 提出划分的原则; 运用数据挖掘工具, 将用户数据输入到聚类算法模型中, 对聚类的结果进行分析, 分析每组客户的特征, 分析可能流失的原因; 针对选择的目标客户群, 研究保持客户的营销手段与政策, 是确保客户保持营销活动成功的关键.

客户保持的关键是防止客户流失, 对潜在流失客户进行客户预警. 客户预警主要是对电信客户所处状态的一种判断, 其本质是一种分类问题, 即将现

① 基金项目:湖南省教育厅科学研究项目(11C0724,11C0723)
收稿时间:2012-03-20;收到修改稿时间:2012-05-01

有客户分为两类, 有离网倾向的客户和无离网倾向的客户.

预测客户流失的方法: 应用软件对数据进行挖掘测试, 包括统一的电信客户资料, 客户属性, 购买信息, 消费结构, 模型参数, 模型等; 应用数据分析方法和数据挖掘技术对电信客户流失前的行为进行分析和预测; 应用聚类算法 DBSCAN 和决策树 ID3 的方法对模型应用的实验结果进行过程分析.

结合客户细分模型与客户流失预测模型进行实验, 能够获取客户的流失倾向及业务特征, 方便营销部门选择流失倾向较高的客户, 采取针对性的客户保持策略.

3 电信客户保持方法

随着电信行业中的竞争愈来愈激烈, 获得新客户的开支愈来愈大, 而保持客户比获取新客户更加节约成本, 如何使用数据挖掘技术对保留客户的活动进行建模, 对整个客户保持工作起着重要的作用. 应用数据挖掘技术进行客户流失分析, 主要是预测哪些是潜在流失客户, 同时评估出最有效的客户保持方法.

常用的数据挖掘方法有: 关联分析; 序列模式分析; 分类分析; 聚类分析; 孤立点分析等[3].

聚类分析是数据挖掘技术中最常用的一种方法, 聚类分析算法主要有划分法、层次法、基于密度法等[4]. 基于密度的方法的代表算法主要有 DBSCAN 算法、OPTICS 算法等[5]. 基于聚类分析的客户关系管理(CRM)可以最大限度地获取客户, 保持客户和进行客户分类, 推行针对性的营销政策, 赢得客户, 从而攫取市场分额和提升盈利能力, 提升企业竞争优势.

决策树是用样本的属性作为结点, 用属性的取值作为分支的树结构[6]. 决策树的根结点是所有样本中信息量最大的属性. 树的中间结点是以该结点为根的子树所包含的样本子集中信息量最大的属性, 决策树的叶结点是样本的类别值. ID3 算法是决策树算法的一种, 是由 Quinlan 首先提出的. 该算法是以信息论为基础, 以信息熵和信息增益度为衡量标准, 从而实现对数据的归纳分类.

本文采用基于密度的 DBSCAN 算法, 结合决策树 ID3 算法, 实现电信进行客户保持的分析.

4 算法在电信客户保持中的应用

4.1 算法思想

ID3 算法在选择分裂属性时, 往往偏向于选择取值较多的属性[7], 然而在很多情况下取值较多的属性并不总是最重要的属性, 这会造成生成的决策树的预测结果与实际偏离较大, 而且运算效率不高, 针对这一弊端, 本文提出将 DBSCAN 算法与决策树 ID3 算法相结合, 利用 DBSCAN 算法对流失客户群数据进行划分, 分析客户流失原因, 结合决策树算法中的 ID3 算法用来进行客户流失预测.

首先将流失客户群分类, 从测试数据中抽选出一个未处理的点, 如果抽出的点是核心点 , 那么找出所有从该点密度可达的对象, 形成一个簇; 否则, 抽出的点就是边缘点(非核心对象), 跳出本次循环, 寻找下一个点; 直到所有的点都被处理完成.

然后进行客户流失的预测, 利用决策树 ID3 算法将分类结果实现客户流失预测. 利用 DBSCAN 算法进行分类的基础上, 用 ID3 算法构建客户流失决策树, 然后进行客户流失预测, 实现客户流失的预警分析, 找出客户流失的特征, 帮助电信公司有针对性地改善客户关系, 避免客户流失.

4.2 算法步骤

利用 DBSCAN 算法进行流失客户群分类, 具体步骤如下:

(1) 检测数据中尚未检查过的对象 p, 如果 p 未被处理(归为某个簇或者标记为噪声), 则检查其邻域, 若不小于包含的对象数, 则建立新簇 C, 将其中所有的点加入簇 C;

(2) 对 C 中所有尚未被处理的对象 q, 检查其邻域, 若包含了对象, 则将未归入任何一个簇的对象加入簇 C;

(3) 重复步骤(2), 继续检查 C 中未处理的对象, 直到没有新的对象加入当前簇 C;

(4) 重复步骤(1)~(3), 直到所有对象都归入了某个簇或标记为噪声.

该算法采用欧几里得距离来衡量某个样本点是属于哪个类簇. 当所有的类簇的质点都不再发生变化时, 聚类结束.

欧几里得距离定义: 欧几里得距离(Euclidean distance)也称欧式距离, 在 m 维空间中两个点之间的真实距离. 欧式距离的公式:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})}$$

令: $d_{ij} = d(x_i, x_j), D = (d_{ij})$ 形成一个距离矩:

$$\begin{bmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \vdots & \vdots & & \vdots \\ d_{n_1 1} & d_{n_1 2} & \cdots & 0 \end{bmatrix}$$, 其中 $d_{ij} = d_{ji}$

利用决策树方法 ID3 算法进行客户流失分析, 具体步骤如下:

(1) 客户数据预处理, 对 DBSCAN 算法进行电信客户分类后的数据属性进行处理.

(2) 用 ID3 算法构建客户流失决策树

在数据预处理后, 进行归纳决策树. 因为 ID3 算法是一个递归的过程.

信息熵的计算: $I(U) = -\sum_i P(u_i)\log_2 P(u_i)$

条件熵计算:

$$I(U/V) = -\sum_j P(v_j)\sum_i P(u_i/v_j)\log_2 P(u_i/v_j)$$

信息增益计算: $Gain(X,S) = I(U) - I(U/V)$

(3) 客户流失预测

对电信客户数据样本进行分析, 获得不同属性上的信息增益, 生成决策树, 并转换成一个 If-Then 规则的集合, 生成规则和决策树, 然后对新数据进行分析和预测.

5 测试数据及运行结果分析

测试数据集是某电信公司的客户信息数据库, 数据量为最近四个月的客户相关数据.

实验的硬件环境: PC 计算机, CPU 为 PⅣ2.0G, 内存为 2G; 软件环境: Window XP;

编程环境: Eclipse3.3/myeclipse6.0GA, tomcat6.0.

5.1 聚类结果分析

本文采用 DBSCAN 聚类方法对各类特征分量进行分类, 以确定流失用户在客户价值区间、自然属性、地域区间等各种特征分量的分布特性, 以得到流失用户的共性特征, 并结合领域知识经验, 获取决策树 ID3 的生成规则, 指导决策树的生成. 将 ID3 算法应用于客户流失分析, 能够帮助电信公司深入了解客户流失的原因, 改进客户服务, 对提高客户的留存率具有十分重要的应用价值. 利用DBSCAN算法将客户分为有离网倾向客户和无离网倾向客户, 聚类结果如表 1 所示.

表 1 各月数据分类情况

月份	有离网倾向客户	无离网倾向客户
9月	10680	22469
10月	13787	21861
11月	10293	21765
12月	12579	21340

5.2 性能分析

两种算法相结合, 能够适应海量的电信客户数据分析, 结合决策树方法进行测试, 预测准确率比较如表 2, 表 3.

表 2 决策树 ID3 算法的预测准确率

月份	正确预测用户数	错误预测用户数	决策树 ID3 算法预测准确率
9月	3825	937	80.32%
10月	2057	786	72.35%
11月	1267	585	68.41%
12月	2130	663	76.26%

表 3 DBSCAN/ID3 算法的预测准确率

月份	正确预测用户数	错误预测用户数	结合两种算法预测准确率
9月	3625	637	85.05%
10月	1857	486	79.26%
11月	1467	385	79.21%
12月	1230	263	82.38%

本文在采用 DBSCAN 算法进行电信客户分类的基础上, 针对细分后的客户群, 采用决策树 ID3 算法进行客户流失预测, 从而实现客户保持的研究.

实验结果表明, 两种算法的结合使用, 使得客户流失预测准确率上得到较大提高, 对电信客户的流失预测起到了一定的作用. 结合使用两种算法进行客户流失的分析和预测是可行的和有效的, 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流失率高的客户做好服务工作, 防止客户的流失, 这对于提高电信企业的竞争力、改善客户关系具有重要意义.

参考文献

1 Han JW, Kambr M. Data Mining Concepts and Techniques. Beijing: Higher Education Press, 2001.

（下转第 203 页）