数据挖掘在常旅客系统中的应用

陈秋明 （南京航空航天大学）

摘要：本文通过数据清洗和检查，数据整合，数据构造和数据整合等步骤进行数据处理，采用决策树分类方法进行分析，建立了一套对航空公司常旅客进行评价和监控的标准和方法。

关键词：常旅客系统 数据挖掘 决策树

1 引言

常旅客计划是航空公司争取市场份额，培养忠诚旅客群的有效市场策略，常旅客计划的航空公司进行收益管理的基础，而常旅客的飞行数据也是完善公司航线网络建设、制定航班计划的提供有效的数据支持。因此分析常旅客系统，来提高旅客的忠诚度，是常旅客系统要解决的核心问题。

根据 IDC 公司对全球前一百家航空公司的高层管理人员就什么是航空公司未来战略武器的问题的调查，58% 的人认为是改进对客户的服务；56% 的人认为是增加客户的忠诚度；44% 的人认为是增加市场份额；43% 的人认为是优化联盟伙伴[4]。从数据不难看出，未来的航空市场竞争中，客户服务和提高客户忠诚度将起到相当重要的作用，取得客户的忠诚度是航空公司在竞争中取得优势的源泉。

2 运用数据挖掘提高忠诚度

2.1 数据挖掘的应用现状

常旅客的竞争就是信息的竞争，建立常旅客系统的目的就是为了提高航空公司的搜集常旅客信息的能力，其本质就是提高企业的判断能力，即能够通过已有的事实判断那些未知因素具有的价值，以及对公司未来发展会产生的影响。我们可以对常旅客的信息进行收集，并对收集到的信息整理分类和评估，挖掘其潜在的价值，利用常旅客的系统准确统计常旅客的数量，计算常旅客贡献的利润。识别高价值的旅客，针对性地设计产品，从而提高旅客忠诚度。

2.1.1 客户细分

航空公司的常旅客数量庞大，由于常旅客的个人背景和习惯不同，以及各种因素的影响下，形成了不同的消费特性。在航空公司为其常旅客服务时，如果只提供一样的服务，既不能使常旅客得到满意的服务，也会浪费航空公司的服务成本。因此航空公司需要把常旅客进行细分，把常旅客按照不同的细分类型的提供不同的服务，使旅客获得个性化的服务，从而提高旅客的忠诚度。这样分类的要求是每一个类中的常旅客的属性尽量相同，而类之间的差别越大越好。

从数据挖掘的知识可以得知，数据挖掘中的分类和聚类可以实现目的。在常旅客的客户细分中，聚类是比较适合的，最主要的原因是在开始分类之前，并不知道如何分类和分成几类。而聚类恰恰是这样工具，通过聚类，可以识别数据中的密集和稀疏的区域，从中发现分布模式，以及数据属性之间的相互联系，聚类是一种无指导的观察性学习，也就是说聚类会根据数据本身的特征形成的簇，在没有特定的规则条件下，聚类把整个数据库分成不同的群组，它的目的就是要群与群之间差别很明显，而同一个群之间的数据尽量相似。这显然和常旅客的客户细分的目的是吻合的。

聚类就是将数据划分或分割成相交或不相交的群组的过程。通过确定数据之间在预先指定的属性上的相似性就可以完成聚类任务，最相似的数据聚集成簇。由于簇不是事先定义的，因此在聚集之后要有一个对业务很熟悉的人来解释这样分群的意义。很多情况下一次聚类你得到的分群对你的业务来说可能并不好，这时你需要删除或增加变量以影响分群的方式，经过几次反复之后才能最终得到一个理想的结果。神经元网络和 K－均值是比较常用的聚类算法。K－均值是一种迭代的聚类算法，迭代过程中不断地

移动簇集中的成员直至得到理想的簇集为止。利用K－均值聚类算法得到的簇，簇中成员的相似度很高，同时不同簇中成员之间的相异度也很高。给定簇$K_l=\{t_{l1},t_{l2},\cdots,t_{lm}\}$，则其均值定义为：

$$m_l=\frac{1}{m}\sum_{i=1}^{m}t_{li}\quad\text{（公式 2－1）}$$

K－均值聚类算法的工作步骤可以描述如下：

（1）随机选取K个记录，作为种子节点；

（2）对剩余的记录集合，计算每个记录与K个种子节点的距离，将每个记录归到最近的那个种子节点，这样整个记录集初次划分为K个聚集；

（3）对每个聚集，计算聚集的质心（聚集中心点）；

（4）以每个质心为种子节点，重复上述步骤，直至聚集不再改变。

K－均值算法的时间复杂性为O(tkn)，其中t是迭代次数。K－均值找到的是局部最优解，而不是全局最优解。K的典型取值是2～10。

通过以上算法，可以在常旅客系统对旅客细分。

2.1.2　关联销售

关联销售是指通过分析客户的所购买产品之间的联系，来分析顾客的购买习惯。在航空公司的常旅客系统中，根据常旅客的购买行为进行关联销售行为的分析，是十分必须的，例如在常旅客中需要进行某条航线的推广和促销，就需要在常旅客系统中寻找乘坐这条航线的旅客，还同时会乘坐哪些航线。这样就可以在旅客中把这些航线进行组合，进行促销。而且产品的目标群也非常明确。这种促销的效果要远远好于不分对象的全员促销，既节约了促销成本，又能得到更大的收益。而且类似这样的需求，在常旅客系统中的应用是非常广泛的。

根据这样的目的，可以用数据挖掘中关联规则来实现，首先我们来阐述一下关联的关则的原理，在后面的章节中，我们会给出具体的例子。

关联规则是形如X = = >Y的表达式，其中X和Y是项的集合，这个规则的含义是包含X的事物通常也包含Y，它可以数据库中寻找重复出现概率高的模式，可以展示数据项间未知的关联关系。如乘坐了A航线和B航线的旅客有85%同时也乘坐了C航线，用规则表示为A,B = = > C(85%)。

关联分析可数学形式化地描述为：

定义1 设$I=\{i_1,i_2,\ldots,i_m\}$是由m个不同的属性（谓词或项目）组成的集合（习惯上我们还称I为项集，这里项集中的元素可能是谓词或项目，给定一个数据库D，其中的每一个记录T是I中一组属性的集合：即T包含于I。若集合X包含于I且X包含于T，则记录T包含集合X。一条关联规则就是形如X = = > Y的蕴涵式，其中X包含于I，Y包含于I，$X\cap Y=\varnothing$

关联规则X = = > Y成立的条件是，

（1）它具有支持度S。即在数据库D中至少有S%的记录包含X U Y；

（2）它具有置信度C。即在数据库D中包含的X记录至少有C%的同时也包含Y。习惯上将关联规则表示为X = = >Y（S%，C%）。

支持度：Support% = The Number of Transactions (X U Y) /The Number of Transactions(D)

置信度：Confidence% = The Number of Transactions (X U Y) /The Number of Transactions(X)

· 其中：支持度定义了项目在整个数据库中所占的比例；置信度定义了发现规则的强度，有时置信度也被称为正确率，支持度也被称为覆盖率。

在关联规则中有很多的算法，其中Apriori算法载效率和优化等方面都表现的比较好，通过计算，可以得到很多X = = >Y条规则，而如何使用则要看具体的业务应用了。

2.1.3　忠诚度的分析

对于常旅客的忠诚度的分析的目的，是要通过对忠诚旅客的特征的了解，从而能够做到常旅客的忠诚度进行预测分析。可以通过数据挖据中的预测方法来实现。整个忠诚度模型的建立是在对于常旅客忠诚度变化的一种预测，属于预测类型的知识。由于决策树在数据挖掘的关键的特性上都获得了较高的评价，因此它可以在多种多样的商业问题中用于探究和预测。于建立模型和要发现的交互作用比现实生活中的问题时，正是决策树所擅长的，而且利用决策树的方法得到的模型比较容易被人们理解，所以我们在建立常旅客的忠诚度的模型时选择使用决策树。

下面我们主要介绍一个决策树的算法实现。

生成决策树的过程通常分为两个阶段：建树

(Tree Building)和剪枝(Tree Pruning)。建立决策树的基本算法是贪心算法,它以自顶向下递归的各个击破方式构造决策树,其中又以 Quiulan 研究的 ID3 算法影响最大。剪枝的目的是降低由于训练集存在噪声而产生的起伏。

第一步:建树

建树是决策树分类的基础,下面以传统的工 D3 算法来说明建立决策树的过程。决策树归纳学习算法以 ID3 为代表,ID3 学习算法采用分治策略,在决策树的递归构造过程中,在树的节点上利用特征的信息增益大小作为分枝属性选择的启发式函数,选择信息增益最大的特征作为分枝属性。ID3 具有描述简单,分类速度快的优点,适合于大规模数据的处理,下面是 ID3 的算法描述。

ID3 算法的形式化描述

输入:训练样本 samples,由离散值属性表示;候选属性的集合

attribute list

输出:一棵判定树

第二步,剪枝。

当决策树创建时,由于数据中的噪声和孤立点,许多分枝反映的是训练数据中的异常。为了保证决策树分类的质量和精确度,必须要能够去除数据中的噪声,修剪所构造的决策树。通常使用统计度量,剪去最不可靠的分枝,这将导致较快的分类,提高树独立于测试数据正确分类的能力。主要有两类剪枝方法:

(1) 向前剪枝(pre – pruning)。在建树的过程中,当满足一定条件,例如信息增益 Information Gain 或者某些有效统计量达到某个预先设定的阈值时,节点不再继续分裂,内部节点成为一个叶子节点。叶子节点取子集中频率最大的类作为自己的标识,或者可能仅仅存储这些实例的概率分布函数。然而,选取一个适当的阈值是困难的,因为较高的阈值可能导致过分简化的树,而较低的阈值可能使得树的化简太少。

(2) 向后剪枝(pos – pruning)。与建树时的训练集独立的测试数据进人决策树并到达叶节点时,测试数据的类别与叶子节点的类别不同,这时称为发生了分类错误。当树建好之后,对每个内部节点,算法通过每个枝条的出错率进行加权平均,计算如果不剪枝该节点的错误率。如果裁减能够降低错误率,那么该节点的所有儿子就被剪掉,而该节点成为一片叶子。出错率用与训练集数据独立的测试数据校验。最终形成一棵错误率尽可能小的决策树。向后剪枝所需的计算比向前剪枝多,但通常产生更可靠的树。在实际应用中可以交叉使用向前剪枝和向后剪枝,形成组合式方法。

显然在数据挖据在常旅客系统中的应用,远不止这些,还可以进行促销政策、主常旅客的流失率的分析等,如果航空企业能够降低常旅客的流失率,收益也非常可观的,因为流失率每降低 5%,利润有可能增加 25 ~ 85%,如此种种,由此可见,在常旅客系统,数据挖掘是大有作为的。

3 结论

本文主要阐述了在民航业的常旅客系统中,如何通过数据挖掘的方法,使用忠诚度的指标对常旅客的忠诚的变化情况做出预测分析,使得航空公司在此基础上提前采取措施,顺利的实施常旅客的忠诚度计划。通过对常旅客的忠诚度的建模分析,使航空公司对于常旅客的整体行为有了比较全面的了解,也可以使公司针对常旅客的特征,细分市场,对于不同的旅客给与不同的服务,从而在降低服务成本的同时,获得较好的利润。

参考文献

1 Nigel Hill & Jim Alexander 客户满意度和忠诚度的测评手册,机械工业出版社, 2004 –3.

2 Merlin Stone & Bryan Foss 卓越的客户关系管理,华夏出版社, 2003 –9.

3 基于旅行价值链的信息系统建设,空管在线, 2004 –3.

4 在 CRM 中有效利用数据挖掘,孙波, 2004 –6.

5 忠诚计划与成功营销,刘莉、孙卓, 2004.

6 李天华,常旅客管理之我见,空运商务, 2005 –3.