**《大数据分析技术综合设计》**

**题目：A航空公司客户流失分析**

在国内航空市场竞争日益激烈的背景下，A航空公司在客户流失方面应该引起足够的重视。如何改善流失问题，继而提高客户满意度、忠诚度是航空公司维护自身市场并面对激烈竞争的一件大事，客户流失分析将成为帮助A航空公司开展持续改进活动的指南。（教材215页）

客户流失分析可以针对目前老客户进行分类预测。针对航空公司客户信息数据（见demo/data/air\_data.csv），可以进行老客户以及客户类型的定义（其中将飞行次数大于6次的客户定义为老客户，已流失客户定义为：第二年飞行次数与第一年飞行次数比例小于50%的客户；准流失客户定义为：第二年飞行次数与第一年飞行次数比例在区间[50%,90%)内的客户；未流失客户定义为：第二年飞行次数与第一年飞行次数比例大于90%的客户）。同时需要选取客户信息中的关键属性如：会员卡级别、客户类型（流失、准流失、未流失）、平均乘机时间间隔、平均折扣率、积分兑换次数、非乘机积分总和、单位里程票价、单位里程积分等。随机选取数据的80%作为分类的训练样本，剩余的20%作为测试样本。构建客户的流失模型，运用模型预测未来客户的类别归属（未流失、准流失，或已流失）。

具体要求如下：

1. 根据要求，对三类客户进行筛选和定义；
2. 按照8：2的比例，随机划分训练数据和测试数据，并将训练数据和测试数据分别写入mysql数据库；
3. 分别从mysql数据库中读取训练数据和测试数据，并进行标准化；
4. 分别采用决策树，支持向量机，BP神经网络三种方法预测未来客户的类别归属，对预测结果进行评估，并对三种方法的结果进行比较分析。

* 实战背景
  + 如今的营销已经远远不同于过去了，那个一张传单一则广告的时代结束了，这是个大数据的时代。
  + 一个时代有一个时代的生存法则，那些没有适应时代的企业，即使曾经是个庞然大物，今天也几乎不见踪影。
  + 在航空公司这个特殊的领域，国内竞争还是很严峻的，****一个顾客的流失造成的损失是4-5个新顾客的流入所不能弥补的****，分析用户（特别是会员用户）相关信息，建立模型，发现流失用户特征，制定针对性营销策略，挽留用户是企业生存的重要一环。
  + 本项目基于大数据样本进行挖掘建模，根据用户特征建立模型发现用户流失特征及其原因分析。
* 分析
  + 其实，作为航空公司，客户流失是不可避免的，因为有些用户乘机频率非常低，这部分用户是偶然用户，他们的流失没有太大的挽留价值。
  + 再者，那些会员用户才是研究的重点，愿意入会的，一定是有频繁乘机体验的，这部分人要么是商务精英要么是比较富裕的人，这几类人都是颇具头脑的。但是，即使这样还是流失了，那就是航空公司给出的服务问题了，这类人是重点研究对象。
* 处理过程

当前，有一个比较尴尬的问题，什么样的顾客可以定义为流失呢？样本数据中根本没有这一项。我们不妨这样定义：

| **客户类型** | **定义方式** |
| --- | --- |
| 已流失客户 | 第二年飞行次数与第一年飞行次数比例小于50% |
| 准流失客户 | 第二年飞行次数与第一年飞行次数比例在[50%,90%)内 |
| 未流失客户 | 第二年飞行次数与第一年飞行次数比例大于90% |

通过上面的定义，解决了这个问题，以一种相对合理的方式定义了用户类型。

当然，实际处理时还需要选取客户的关键属性建立模型，如会员卡级别、客户类型（如上定义）、平均乘机时间间隔、平均折扣率、积分兑换次数、非乘机积分总和、单位里程票价和单位里程积分等。

任然采用数据挖掘领域常用的82开进行训练测试样本分布。

数据获取

数据给出（特征过多）

数据探索

首先声明，表头均没有使用中文，均为航空公司报表专业术语，观测窗口使用两年（06年4月到08年3月），请了解后再阅读本博客，这里不做赘述了。

经过初步的数据探索，我得到了一个分析表，记录数目，各列的平均值、四分位数、标准差、极值等；

为数据预处理定下了基本的基调，首先，有几列数据几乎所有记录都为0，这是没有意义的。（在数据挖掘领域中没有对比的数据是没有任何意义的，因为所有数据都是为了建模服务的）

数据预处理

在数据预处理的第一步就出现了一个不算问题的问题，就是由于Excel解析csv以及编码的一些问题，我看到的表格是错位的，我数据处理第一步就直接数据流导出去成为一个新的xls文件，对于xls文件Excel的解析没有问题，现在，最基本的，我们可视化判断没有了阻碍。（保留最开始的数据是数据分析的基本原则之一）

然后，发现第一列用户编号是用户的唯一标识，然而这对于建模是没有意义的，必须剔除第一列。

选取合适属性构造了需要的六个属性LRFMCK模型，并进行标准化数据。

数据挖掘建模

流失用户的分析，并且判断用户会不会流失，这是一个分类问题，我们使用决策树建模。

后续处理

通过分类预测现有用户是否流失倾向是无比重要的，发现流失可能性进行相应营销比较必要，这里决策树预测准确率还算可以，在深度合适（发现4-5）预测准确率达到85%左右，对于营销需求是达到了。