客户流失预测模型设计与实现①

Design and Implementation of Customers Churn Prediction Model in Telecommunication

周生宝　郭俊芳　(大同大学 数学与计算机科学学院 山西 大同 037009)

摘　要： 有效地防止用户流失，降低流失率成为目前各个电信运营商急需解决的难题。本模型依据数据挖掘原理，以 CRISP_DM(Cross-industry Standard Process for Data Mining)建模过程为框架，以与某地联通公司合作的“电信客户流失预测系统”项目为依托，利用挖掘软件 SPSS Clementine 8.0 逐步建模，实现了电信行业客户流失预测系统。

关键词： 客户流失预测 CRISP_DM 多基决策树联合决策 Clementine C5.0 算法

在电信业，获得新客户的费用是较高的，而挽留一个客户有时可能仅需一个电话，因此怎样前瞻性地发现有流失倾向的客户并成功挽留是各大电信运营商需解决的问题。本文模型采用 SPSS 公司的 Clementine 8.0[1]作为平台，采用数据挖掘技术，严格按照 CRISP-DM(Cross-industry Standard Process for Data Mining)[2]行业标准逐步以商业理解，数据理解，数据准备，建立模型，模型评估与发布的步骤来建立模型，预测在网客户未来一段时间内流失可能性以及流失原因，为客户挽留提供依据。

1　数据处理

1.1 商业理解

客户流失本质是一种分类问题，即将现有客户分为两类：有流失倾向的客户和无流失倾向的客户。模型的输入变量选择两类数据：静态数据和动态数据。静态数据是客户自然属性，如性别、年龄、收入、婚姻状况、学历、职业、居住地区等等。动态数据是经常或定期改变的数据，如每月通话金额、交费纪录、消费特征等等。模型预测的目标变量是流失与否这个状态。

1.2 数据理解

模型用 Clementine 中的表格节点(table)、数据查看节点(data audit)、质量节点(quality)、统计节点(statistics)等查看客户情况。也可用直方图(histogram)、分布图(distribution)初步确定哪些因素可能影响客户流失，之后在数据准备中对数据进行处理。

1.3 数据准备

数据准备包括对数据的选择、清洗、缺失值的处理、属性转化、衍生变量的生成、离散化、抽样等等[3]。系统用 filter 节点滤掉只有一个值的属性和与流失无关的属性，比如 query_kind(查询方式)，pay_s_id(帐户销帐方案)等属性与预测结果无关可去掉。属性转化是指由已有的属性生成其他未知的属性，如用户自然属性表中没有年龄属性和性别属性，但由相关专家知识知道这两个属性与流失可能性有关联，可以用相应公式根据用户填写的身份证号生成。系统用生成节点生成了身份证号为 15 位的纪录的性别，生成公式设置为：to_integer(substring(15,1,to_string(IDENTITY_CODE)))mod，即取出身份证号最后一位数除以 2 取余数，最后一位为偶数的，余数为 0，这种情况记录为女性；最后一位为奇数的，余数为 1，记录为男性。对于顾客填写的明显错误的身份证号导出的结果值需用均值、默认值代替错误值或根据现有正确数据的分布比例导出替代值，本模型用 filler 节点重新填写错误年龄数

① 基金项目:大同大学 2008 年度青年科研基金项目(2008Q15)
收稿时间:2008-11-25

据。当用直方图节点得到年龄分布图后，系统用 derive 节点把年龄离散化为 1、2、3、4、5 五个值。衍生变量是用来抓住在用户流失前有异常变化，变化中隐含了用户的行为信息的一些属性。大多数通话行为属性属于这类，如正常情况下用户月市话量基本稳定，但是很多流失用户在流失前这个值就会减少或增大。针对这些变化形式模型设计了平均值和波动值 2 类衍生指标。图 1 为数据处理模型图。

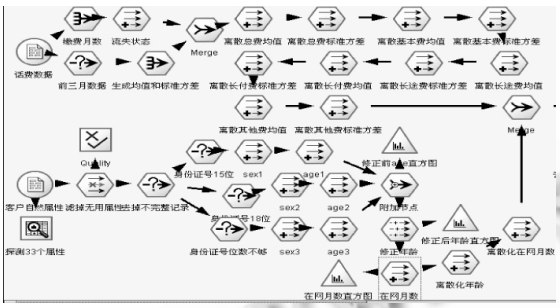

图 1　数据处理模型图

2　建立模型

系统用多基决策树联合决策思想，它模仿现实生活中多个专家共同参与一个问题的决策，这些专家每个人的意见都有瑕疵，最后由另外一个专家综合他们的意见形成一个最终决策[4]。经实验验证采用多基决策树联合决策算法比用一个决策树预测的准确率多数情况要好，因为它组合了多个分类器，提高了分类器的泛化能力[5]。

模型采用的具体算法为：从整理好的流失数据与未流失数据中随机选出 1/3 数据形成训练集，其余数据为测试集。把测试集中的未流失客户数据等分成 n 份，每一份与测试集中的流失数据合并在一起形成训练子集。并分别用 C5.0 算法生成基决策树。之后采用基于权重的投票策略对测试集中每个客户进行最终决策。设预测类别为流失的权重值为 W1,预测为未流失的权重值为 W2(W1>W2)，n 个基决策树预测类别为流失的置信度之和为 C1，预测类别为未流失的置信度之和为 C2。若，则最终预测结果为不流失，反之则预测最终结果为流失。

我们分析某地联通公司 2006 年 1~6 月的数据，从中随机抽取流失者数据 6657 份，未流失者数据 60000 份。其中未流失数据随机抽取 40000 个记录和流失数据随机抽取 4000 个记录作为训练集，未流失数据中剩余的 20000 个记录和流失数据中剩余的 2657 个记录作为测试集。测试集中的流失者约占 11.727%，比例分布比较符合实际情况，这样预测结果的有效性好。之后训练集中的未流失数据再随机分成 4 等份，每一份和 4000 个流失数据以 2.5：1 的比例合并为一个训练集。用 Clementine 工具中的 C5.0 算法生成 4 个基决策树对测试集进行预测，对每个客户都生成预测状态值$C-state 和置信度值$CC-state。经过多次试验后基决策树的各个参数设置为： number of trials(弱规则数)设为 5，剪枝度设定为 25%，流失客户预测为未流失客户的误分类成本设置为 1.6，未流失客户预测为流失的误分类成本设置为 1，图 2 为生成 4 个基决策树模型图。

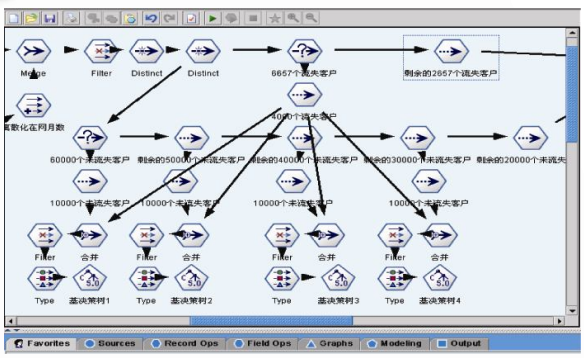

图 2　4 个基决策树模型图

用生成节点生成一个$CC-churn 字段，其意为计算 4 个基决策树预测为流失的加权总置信度，其值为 4 个基决策树中预测值为 1 的置信度之和与权值 1.2 的乘积。再生成一个$CC-nochurn 字段，其意为计算 4 个基决策树预测为未流失的总置信度，其值为 4 个基决策树中预测值为 0 的置信度之和。如果 '$CC-churn' >= '$CC-nochurn'，则预测为流失，$C-state=1。否则预测为未流失，$C-state=0。图 3 为基决策树联合决策的模型图。图 1、图 2、图 3 连接起来就是一个完整的电信业流失预测模型图。

3　模型评估及发布

我们用预测查准率、查全率、提升率评估模型对流失者的准确预测能力。预测查准率是预测流失中实际流失的比例，体现了模型对流失客户的预测是否精确。预测查全率是实际流失中预测正确的比例，体现了模型预测结果的覆盖程度。提升率衡量预测模型捕

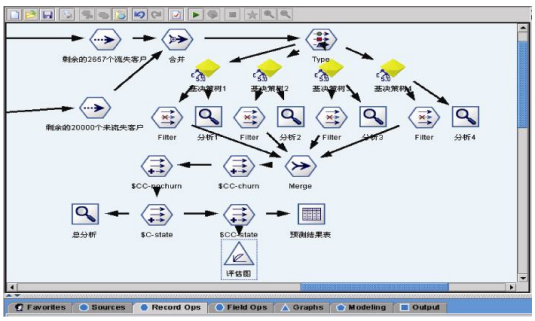

图 3　基决策树联合决策模型图

捉流失客户的难易程度。这三个指标值越大，说明模型预测效果好[6]。Clementine 中的 analysis 节点对 4 个基决策树和联合决策预测结果生成了分析值,图 4 为联合决策后对测试集的预测分析图。综合计算这 5

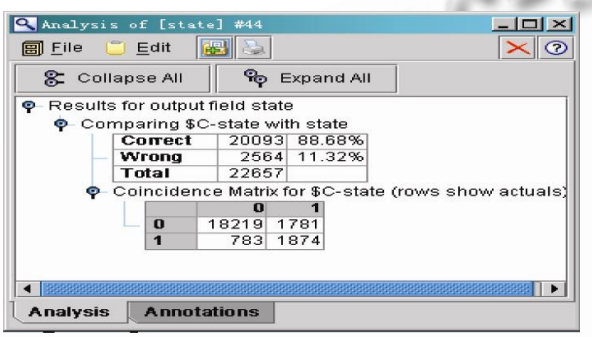

图 4　联合决策后对测试集的预测分析图

表 1　4 个基决策树和联合决策预测结果指标值

算法	准确率(%)	查准率(%)	查全率(%)	提升率(%)
基决策树 1	71.64	22.81	59.50	1.95
基决策树 2	79.71	32.22	66.20	2.75
基决策树 3	75.20	28.80	75.69	2.46
基决策树 4	80.56	32.92	63.38	2.81
联合决策	88.69	51.27	70.53	4.37

个 analysis 节点的值生成表 1。从实验数据可看出联合决策的流失预测效果要好于各个基决策树的预测效果。

模型建好后使用 Clementine 的发布节点(pub－lisher)将流嵌入到运营商自己外部的应用软件中。

4　结论

模型依托某地联通公司的客户数据用 Clement－ine 工具建立，改进了决策树中的 C5.0 算法,采用多基决策树联合决策的算法。通过变化事例空间，增大样本中流失数据所占比例，构造多个分类器，解决了流失客户与未流失客户的比例倾斜问题，而且比只用 C5.0 算法建模更有效，提高了预测精度、普适性和泛化能力。

参考文献

1 SPSS .Clementine User Guide.2003.

2 CRISP-DM 协会.CRISP-DM1.0 数据挖掘方法论指南, 2000.

3 童凤茹.基于组合分类器的信用卡欺诈识别研究.计算机与信息技术, 2006,(7):8－12.

4 WEI C-P, CHIU I-T. Turning telecommunica -tions call details to churn prediction: a data mining approach. Expert Systems with Applications, 2002,23(2):103－112.

5 Dietterich TG. Ensemble Methods in Machine Learning: Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science. 2000. Springer-Verlag, 1857(1):1－15.

6 于爱民.利用数据挖掘实现电信行业客户流失分析.广东通信技术, 2004(5):12-14.