对电信公司客户是否会终止服务的预测

周昊一*

(南京大学 计算机科学与技术系, 南京 210093)

Prediction for whether a customer of telco will terminate a service

Haoyi Zhou*

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

Abstract: Maintaining customers and ensuring customers are satisfied with the products offered have always been a major challenge for telcos. Companies often try to reduce the number of customers terminating their services to avoid losing their market share. In this paper, we used some concrete techniques to discover the probability of each customer to churn, and generated a list contains top 5% of existing customers who are most likely to churn.

Key words: data mining, classification, real world

摘 要: 对于电信公司来说,维持客户并保证客户满意所提供的服务一直是一个很大的挑战.公司通常希望通过减少终止服务的用户数量来避免市场份额的损失.在这篇文章中,我们使用了一些具体的技术来计算每个客户终止服务的概率,并且生成了一个包含了前5%的最可能终止服务的客户的名单.

关键词: 数据挖掘, 分类, 实际问题

中图法分类号: TP301 **文献标识码**: A

1 问题分析

此次需要解决的是一个现实中的问题.一个电信公司想要能够判断出某一个客户(小型和中型的公司客户)是否会在将来终止他的某一项服务.

具体地说,我们需要用所给的一系列数据,给出一个包含前 5%最有可能在 2012 年的前三个月终止某一项服务的现存客户的名单.

2 数据预处理

2.1 数据的整理

这次的数据是一种层次状的结构.

首先每个客户有一个唯一的身份标识 CustomerID.在这一个 CustomerID 下面,有一系列的服务,每个服务是以唯一的 ServiceID 作为标识.每个 CustomerID 还对应一个每月份的 Complaint 记录.

* 作者简介:周昊一, 学号 MF1233055。

每一个 ServiceID 还对应了一个 BillingAccount,用于索引费用的相关记录.因此每个服务下总共包含了以下几项:

1. 每个月的语音使用情况,包含市话,长途,国际长途,网络电话以及月账单.
2. 每个月的总的账单,包括上月未付,本月应付,信用优惠以及本月实际应付等.
3. 对于宽带用户,还有每个月的上传及下载的流量.

由于所有的数据分散到了 6 张表之中,无法直接进行分析,因此需要先进行一步数据整理,将数据按照用户及服务进行整理,最终形成一个树状的数据结构.

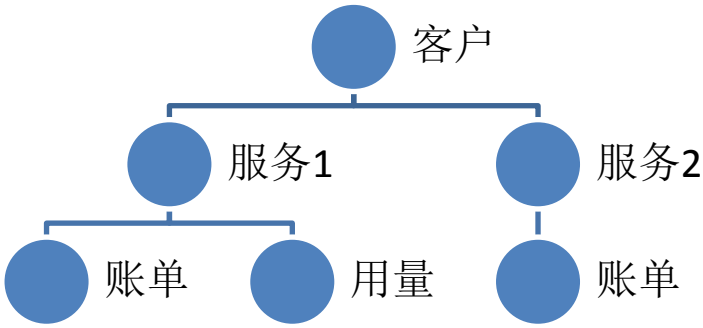

图 1.数据的层次结构

2.2 数据表示

对于这样未经加工过的数据进行处理是很有必要的.首先如果将按照每个用户来算,再将每个服务的每个月的细节都扩展出来,平均每个用户的属性将至少多达上百个,并且由于每个用户所持有的服务数并不相同,每个用户的属性数量还不一样.并且由于服务的个数没有上限,属性的数量都无法确定.

果考虑将一个用户的所有服务取平均呢?这似乎也是不可取的,因为并没有证据显示一个用户在对待每一项服务时的行为都是一致的.并且从观测的角度来说,直接观测到的结果是某项服务的终止.因此从直觉上来说更适合将服务作为分析的主体.

此我们的总体思路就是先按照服务为单位,将数据进行整合.对于每一个服务,都可以分进终止和未终止两类中去.因此我们的训练数据就是在 2011 年中被终止以及至今未被终止的所有服务,我们希望在这样的数据集上训练出来的模型能够预测出一个服务是否会在 2012 年的前三个月中被中止..

此时我们仍然面临着以下几个问题:

1. 如何表示服务在某一个月被终止?
2. 如何对某一个服务进行表示?
3. 从数据中我们可以看到,被终止的服务分布在 2011 年的 1 月到 12 个月的每一个月中.因此很显然,对于像 1 月份就终止的服务,我们缺失了 11 个月的信息,这样的情况该如何处理?

第一个问题和第二个问题是相关联的,一种思路是将服务的 class 定为整型,表示被终结的月份.在预测时只需要预测 class 的值,就可以预测出服务将会被终止的月份.这样的话,对于一个服务的表示就应该是它在 2011 年整年中的所有情况.

仔细思考后会发现,这样做有一个很大的问题,那就是训练集中的 class 分布为 1~12,而需要预测的 class 却均为 13~15(显然不应该是 1~3),因此这样做是不可取的.

因此,只能另辟蹊径,用一个服务在第 K 个月前的 N 个月中的数据对其进行表示.而在预测时,我们只需要预测这个服务在第 K 个月是否会被终止.这样就解决了上面所遇到的预测数据和训练数据无法统一的问题.

这样的话就需要选择一个合适的 K 值.如果 K 太小,所取的数据就无法很好的表示这个服务在最近一段时间的情况,而如果 K 太大,会导致有许多(终结月份<k)的数据变的不完整.同时太大的 K 也无法保证在这个月之后的一两个月内服务仍然没有被终止(比如取 K=12,则相当于所有当前未被终止的服务都被认为是不会在

2012 年一月被终止的).在进行了一些权衡和研究之后,我选择了 6 作为 K 最终的值.这样我们可以保证至少有差不多一半的数据是完整的,并且那些被标为"未终止"的服务在将来的 6 个月内都不会被终止(这样就具有代表性).

2.3 属性选择

在确定了数据的表示方案后,还需要进一步选择所需要的属性.

首先,有许多属性显然和我们需要预测的结果关系不是很密切.比如客户的性别,年龄,以及最早启用服务时间(通常都不是当前服务的启用时间,而是已经终止了的过去的服务),所处的地区等.

这些属性虽然可能和一个服务是否会被终止有某种潜在的联系,但是至少从经验和直觉的角度来看,应该不会太密切,至少不会有直接或者是明显的影响.如果采用了这些数据,可能反而会导致过拟合的情况,因此需要将它们剔除.所剔除的所有的属性为(按照表中的属性名):

CUSTOMER_SINCE,INDUSTRY,T_LOCATION,HSBB_AREA,ACTIVATION_DATE,COMEBACK_DATE, COMEBACK_PRODUCT.

对于剩下的那些属性也不能直接使用,要做一定的处理.

比如像每个月的账单信息,如果直接展开到每一个月,首先有许多项中会缺失某几个月的数据,其次这样会造成属性数量过多,并且属性代表性不强,意义不明确,同样也很可能导致过拟合.

因此我们需要分析一下要能够预测出某个服务在下一个月是否被终止,到底需要什么样的数据.

对于某一个客户而言,终止某一项服务一定是有某个原因的,比如开销过大,效果不好,服务质量差等等.我们虽然无法直接知道这个原因,但它一定会通过某种形式表现出来.

比如当客户对服务不满意时,抱怨的数量就会上升.如果客户是因为不再需要这个服务而终止了它,那么很有可能在前一两个月中这个服务的使用量有明显的下降.

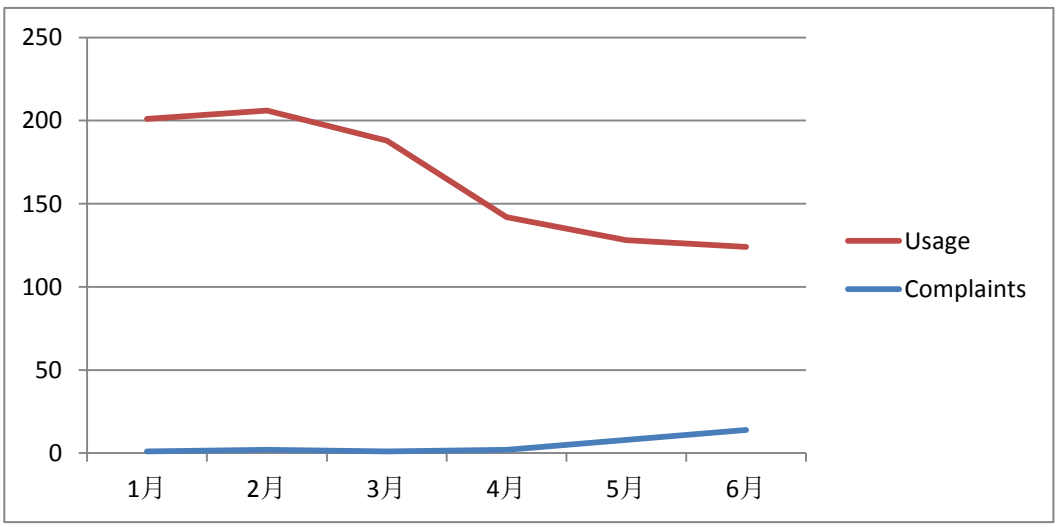

图 2.有可能会终止服务的客户的行为

从上面的分析可以看出,如果用户终止了某一个服务,那么通常来说可以观察到用户近几个月数据中的某种变化.反过来说也是合情合理的:如果观察不到某些特定的变化,那么有很大的可能这项服务不会被终止.

基于以上分析,很明显的,如果只表示出数据的存量是无法很好的区分被终止和未被终止的服务的.我们需要的是表现出数据的流量(至少是通过某种方式间接地表示).

因此很自然的,我们可以通过描述一个服务的长期(long term)和短期(short term)的表现来表示他的近期状

态.长期表现可以用近 K(K=6)个月的数据的平均值来表示,而短期表现可以用近 2 个月的平均值来表示.取 2 个月是因为如果只取 1 个月的话 2012 年 2 月份和 3 份的数据都无法取得,而取 3 个月的话时间跨度又过长(再加上我们并没有考虑被终止的那个月当月的情况).

最后,由于我们现在的数据是以服务为单位的,而有一些属性是属于客户的(比如价格,网速,抱怨的数量等).为了简化情况,这里仅仅是将这些数据复制到每一个对应的服务中,作为服务的属性来看待.

对于同时包含语音和宽带的服务,共有 41 个属性.对于只包含语音的服务,共有 35 个属性.

2.4 数据筛选与生成

由于在数据中需要分别表示出近 6 个月和 2 个月的情况,因此对于某一些数据项来说会有属性缺失的情况.

比如 2011 年 1 月份就终止的服务,显然不太适合作为训练数据,因为前 6 个月的数据全部缺失,甚至前 2 个月的数据也没有,同理 2 月份就被终止的数据也不是很合适.因此在最后的训练集中我将这部分都去除了.

而对于未被终结的服务,则有很多选择的余地,可以任意选择 k~k+5(1<=k<=7)这 6 个月份的数据进行表示.在这里我选择的是 1~6 月份,原因是这样可以保证在将来的 6 个月之内这个服务仍然不会被终结,这样可以保证这个数据对于未被终止的服务具有足够的代表性.

而对于预测用的数据,由于 2012 年 1 月份和 2 月份的数据都是没有的,因此只能用前面几个月的来估计.比如生成 2 月份的预测用数据时,实际上取的分别是 2011 年 8~12 月份的平均(当作 6 个月)和 12 月份的平均(当作 2 个月).而在生成 3 月份的数据时,前两个月的平均值也是用 2011 年 12 月的值来估计.

由于要分别预测 3 个月份的情况,因此对每一个还未被终止的数据都要生成 3 条预测用数据,分别对应 1,2,3 月.

3 分类

从前面的分析可以看出,这个问题可以当作一个典型的分类问题来处理.class 属性的取值范围为{unterminated,terminated}.

3.1 算法选择

由 2.3 节的分析可以看出,一个服务的属性与它所属的分类之间有很明显的逻辑联系.比如从直觉上来说,6 个月内的平均抱怨数较小而近 2 个月的平均值较大,这个服务就很有可能被终止.

因此很明显的,这样一个任务非常适合使用决策树进行归纳.

因此我选择了经典的 C4.5 算法对其进行处理.

经过 10 叠交叉验证,发现准确率(precision)可以达到 70%以上,基本满足要求.

Results

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.984	0.347	0.976	0.984	0.98	0.675	0.903	0.987	unterminated
	0.653	0.016	0.739	0.653	0.694	0.675	0.903	0.679	terminated
Weighted Avg.	0.962	0.325	0.96	0.962	0.961	0.675	0.903	0.966	

图 3.在 voice_only 的训练集上进行 10 叠交叉验证的结果

3.2 后期处理

由于我们最终需要得到的不是所有可能被终止的服务,而是 5%可能终止某个服务的客户.

因此我们还需要对最后得到的分类结果做一下处理.理论上来说对 1 月份的分类应该是最准确的,2 月份的差一点,3 月份的最差(由于数据缺失).但是由于不知道具体差多少,难以设定权重,为了简化问题,还是当成一样的来处理.但是有一点是显然的,即如果一个服务在越多的月中被判为被终止,那它实际被终止的概率肯定就越高.

如果一个客户被判定为会终止一个服务,那么大概有 p 的概率(p 约为 70%)他真的会终止这个服务.很显然,如果这个客户还被判定为会终止另一个服务,那么他终止某一个服务的总概率为$1-(1-p)^2$.也就是属于一个客户的被判定为被终止的服务数量越多,这个客户会终止某一个服务的可能性就越大.

结合以上两点,我们只需要做一个简单的处理,每当有一个服务在某一个月中被判定为被终止,它所属的客户的权重就+1.最后按照权重对客户进行降序排序,取出前 5%.

References:

[1] Jiawei Han, Micheline Kamber 著范明 孟小峰 译 《数据挖掘——概念与技术(第二版)》机械工业出版社.

[2] Freeman, Edmund, and Gabor Melli. "Championing of an LTV Model at LTC." ACM SIGKDD Explorations Newsletter 8.1 (2006): 27-32.