基于Spark深度学习的客户流失分析

温明杰[1,2]，彭 耀[1,2]，李 军[1,2]，任德虎[1,2]，赵致杰[1,2]，张陈斌[1,3]，陈宗海[1,3]
（1. 中科大-象形大数据商业智能联合实验室，安徽合肥，中国，230031；2. 安徽象形信息科技有限公司，安徽合肥，中国，230031；
3. 中国科学技术大学，安徽合肥，中国，230027）

摘 要： 数据挖掘是近年来伴随着人工智能和数据库技术发展而出现的一门新兴技术。基于Spark的大数据挖掘是近年来伴随着人工智能和大数据技术，以及分布式计算、云计算发展而出现的一门新兴技术。它的核心功能是从巨大的、多样性数据集或数据仓库中迅速获取有用信息，以供企业分析和处理各种复杂的数据关系。最近几年，数据挖掘技术以其强大的数据分析功能被普遍应用到电信运营商客户管理之中。在本文中，我们利用数据挖掘实现了电信宽带用户的流失分析，利用Spark的技术实现了并行的聚类算法、决策树算法。

关键词： 数据挖掘；深度学习；决策树；客户流失
中图分类号： TP391

The Loss of Customers Analysis Based on Spark Deep Learning

Wen Ming-jie[1,2]，*Peng Yao*[1,2]，*Li Jun*[1,2]，*Ren De-hu*[1,2]，*Zhao Zhi-jie*[1,2]，*Zhang Chen-bin*[1,3]，*Chen Zong-hai*[1,3]
(1. USTC-ETHINK BigData Business Intelligence Joint Laboratory，Anhui，Hefei，230031；
2. AnHui ETHINK Information Secience&Technology Co.，Ltd，Anhui，Hefei，230031；
3. University of Science and Technology of China，Anhui，Hefei，230027)

Abstract: Data mining in recent years, along with the development of artificial intelligence and database technology and the emergence of a new technology. Spark-based big data mining and artificial intelligence in recent years, accompanied by big data technology, as well as distributed computing, cloud computing and the emergence of a new technology. Its core function is to quickly obtain useful information from the huge diversity of data collection or data warehouse for business analysis and processing of complex data relationships. In recent years, data mining technology for its powerful data analysis capabilities are widely applied to the telecom operator's customer management. In this paper, we use data mining to achieve the churn analysis Telecom broadband users, the use of technology to achieve Spark parallel clustering algorithm, decision tree algorithm.

Key words: Data Mining; Deep Learning; Decision Tree; Loss of Customers

1 引言

在电信运营商重组以及三网融合的大背景之下，基础电信运营商正转变为全业务运营商。全业务运营的核心是通过现有电信业务的有效整合和融合，开展基于宽带和IP化的FMC、ICT以及相应的增值业务。而其中宽带业务的开展以及宽带用户的发展对于运营商发展战略有着重要的现实意义。

2011年初，中国电信高调启动了“宽带中国光网城市”工程，大力发展光纤到户；中国移动加速推进4G建设;中国联通也将从55个城市开始全面推广HSPA+，无线接入带宽提升到20 M。新一轮宽带网建设、宽带业务以及宽带用户发展正全面展开，宽带业务用户具有高ARPU值现金流稳定等优势特点。三大运营商努力发展宽带新兴业务，中国电信明确表示将采取融合策略拓展新增市场，以宽带提速为契机，提高宽带产品渗透比例，将继续坚持/聚焦客户的信息化创新战略，大力拓展移动宽带和行业应用的业务规模，努力做好存量业务保有，持续优化收入结构。

根据相关调查数据显示，用户保有率增加5%将有可能为运营商带来85%的收入增长；挽留成功一位老用户的成本只占发展一位新用户成本的1/5；向老用户推荐新

产品的成功率为向新用户的 3 倍。由此可见，积极采取措施,防止宽带客户流失对于中国电信意义重大。

市场的重要战略意义促使三大运营商着力发展宽带用户。当前，电信市场正从新增市场为主转变为存量市场为主，运营商现有宽带用户的保有工作变得十分必要。目前，三大运营商都面临宽带客户流失问题，客户流失严重影响运营商业务发展以及企业效益。同时，客户流失往往带来的是双重效果：一方面，自身宽带用户减少；另一方面，带来竞争对手运营商的用户增加。如何进行宽带客户流失预警并完成对于预警用户的针对性挽回是目前运营商急需解决的问题。

Billing(计费系统)、经营分析系统、网管系统等这些系统在支撑电信企业运营过程的同时也产生了海量的运营数据，这些数据包括电信用户资料、用户消费情况资料、电信用户账单资料以及电信企业业务发展情况资料等。如何将这些海量的数据变为有价值的信息是电信运营商需要思考的问题。

为解决电信宽带客户流失严重现状，实现将电信面临的宽带客户流失率下降的目标，需要基于 Spark 数据挖掘技术设计并实现一种能够预测出宽带客户流失的有效方法。基于这种方法可以获取电信即将流失的宽带客户名单。对于宽带客户流失名单进行针对性营销，实现宽带客户流失预警及维系挽留，最终达到降低宽带客户流失率的目的。

基于 Spark 的大数据挖掘是近年来伴随着人工智能和大数据技术，以及分布式计算、云计算发展而出现的一门新兴技术。它的核心功能是从巨大的、多样性数据集或数据仓库中迅速获取有用信息，以供企业分析和处理各种复杂的数据关系。最近几年，数据挖掘技术以其强大的数据分析功能被普遍应用到电信运营商客户管理之中。

2 数据挖掘的主要方法

作为一种先进的数据信息处理技术，数据挖掘与传统的数据分析的本质区别在于它是数据关系的一个探索过程，而且多数情况下是在未有任何假设和前提的条件下完成的。数据挖掘具备多种不同的方法，供使用者从不同的纬度对数据展开全面分析。

(1) 相关分析和回归分析。相关分析主要分析变量之间联系的密切程度；回归分析主要基于观测数据与建立变量之间适当的依赖关系。相关分析与回归分析均反映的是数据变量之间的有价值的关联或相关联系，因此两者又可统称为关联分析。

(2) 时间序列分析。时间序列分析与关联分析相似，其目的也是为了挖掘数据之间的内在联系，但不同之处在于时间序列分析侧重于数据在时间先后上的因果关系，这点与关联分析中的平行关系分析有所不同。

(3) 分类与预测分析。分类与预测用于提取描述重要数据类的模型，并运用该模型判断分类新的观测值或者预测未来的数据趋势。

(4) 聚类分析。聚类分析就是将数据对象按照一定的特征组成多个类或者簇，在同一个簇的对象之间有较高的相似度，而不同的簇之间差异则要大很多。从过程上看，聚类分析一定程度上是分类与预测的逆过程。

3 数据挖掘的应用

下面简要地描述数据挖掘在客户流失分析管理中的应用过程。

通过基于行业标准数据挖掘过程模型 CRISP-DM，与业务专家讨论，结合电信业务特点，实现基于 Saprk 的数据挖掘建模。

(1) 定义主题客户流失分析中的主题应当包括流失客户的特征、现有客户的流失概率如何(包括不同细分客户群的流失程度)、哪些因素造成了客户的流失等。主题是数据挖掘的主要目标，决定了此后过程中数据挖掘的主要努力方向，因此在定义上应当十分明确。

经过与业务专家讨论，基于以下业务数据，实现数据挖掘建模。

目标用户定义：

①统计月，用户状态正常；

②入网满三月的非易通卡、非预付费、非掌宽、非无线宽带用户；

③非电信职工、非公免、非公纳及测试用户。

离网定义：

①欠交 3 个月；

②停机保号；

③当月费用≤0；

④三个月合计时长小于或等于 3 小时。

时间窗口定义（对 1 和 4 类用户，离网月为 M=2014 年 6 月，对 2 和 3 类用户离网月为 4 月）。

分析期（即模型的输入变量）：指用户离网前的历史通信行为产生时间段，即模型输入变量的时间窗口（分析期为 M=2014 年 1、2、3 月）。

维系期：指预警名单输出时间，即应用模型预警名单，开展维系工作的时间窗口（维系期为 M=2014 年 4 月）。

反应期（即模型的输出变量）：指离网标识产生时间，即模型的输出变量的时间窗口。

(2) 数据选择。数据选择是数据挖掘的前提，主要是确定数据字段的收集，因为并不是所有的客户信息都会对客户的流失产生影响，应尽可能地降低数据的复杂度

以发掘较高的关联度，但是考虑到后期客户流失的多维分析，应当尽量确保客户信息的完整性，因此，应对客户的有价值信息予以区分收集，剔除部分冗余数据，减少数据噪音。此间要注意的是在客户流失分析上，从数据仓库中采集数据的主要目的是调查客户信息的变化情况，因此对数据采集时间间隔的设置显得尤为重要。若采集时间过长，可能在流失判断出来时客户已然流失；若采集时间过于紧密或者实时采集，则需要考虑运营商现有系统的支撑能力。

表1 部分输入变量

1	SERV ID	客户编号
2	AGE	年龄
3	GENDER	性别
4	ZW MONTH	在网时长
5	BILLING MODE	付费方式
6	LX	来显
	……	……
42	ZB CT13	本月和三月平均长途降幅
43	ZB CT15	本月和五月平均长途降幅

(3) 分析数据。分析数据主要是对提取的数据进行分析，找到对预测输出影响最大的数据字段，并决定是否需要定义导出字段。在分析数据时需要谨慎选择对预测相关的流失客户数据参与建模才能有效建立模型。分析数据过程还应包括数据清洗和数据预处理。数据清洗和预处理是建模前的数据准备工作，主要包括数据抽样、数据转换、缺损数据处理等。数据抽样是根据事先确定的数据进行样本抽取，选择抽样而不是对整体进行处理，以降低系统的处理量。另外样本一般分为建模样本和测试样本，一部分用来建模，另一部分用来对模型进行修正和检验。数据转换是为了保证数据的质量和可用性，比如某些数据挖掘模型需要对连续数据进行离散化、归一化处理等。缺损数据有时可以不做处理，由后面具体选择的数据挖掘模型来处理。

(4) 模型建立。对数据进行分析并利用各种数据挖掘技术和方法在多个可供选择的模型中找出最佳模型。初始阶段可能模型拟合度不高，需要反复更换模型，直到能够找到最合适的模型来描述数据，并从中找到规律。

(5) 模型的评估与检验。模型建立之后，一般要通过训练集的测试才能考虑下一步应用。比较常规的验证方法是输入一些历史的流失客户数据，运行此模式予以判断，比较数据挖掘的结果与已知历史结果的差异。见图1。客户流失判断一般存在两种错误结果。一是弃真错误，即原有历史客户具备流失倾向并且已经流失，但是模型未能够准确预测客户的流失倾向；二是存伪错误，即原有用户并未有流失的倾向，但被模型判断为具有流失倾向。见表2、表3。

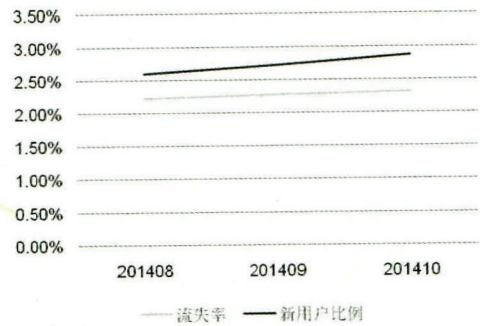

图1 新用户与流失用户对比图

表2 用户表

	总体用户	流失用户		正常用户	
	准确率	1-准确率	1-查全率	0-准确率	0-查全率
公众-单宽	92.66%	40.90%	57.59%	97.30%	94.83%
政企-单宽	92.25%	42.80%	45.33%	96.04%	95.63%
公众-融合	95.76%	44.87%	57.15%	98.31%	97.27%
政企-融合	95.85%	41.72%	48.65%	98.14%	97.55%
校园-单宽	90.82%	35.34%	63.60%	97.55%	92.56%
校园-融合	94.63%	35.05%	62.53%	98.60%	95.80%

表3 聚类表

聚类	符合相应类特点实际离网预测不离网用户数	全量数据符合相应类特点的用户数	符合相应类特点的非离网用户数
聚类-1	1 883	136 166	134 193
聚类-2	951	83 050	82 023
聚类-3	1 883	125 754	7 124 541

说明：

1. 聚类2：漫游费高、短信费高、c网交往圈大、国际交往圈大、通话行为相对于其他两类较高、上网流量高，聚类1和3区别也表现在以上几个属性。

2. 这部分实际离网而预测不离网的用户通话次数、主叫时长、上网流量大部分都很高，这就是为什么被判成不离网的原因。

3. 由符合每一类特点的离网与非离网的用户数统计可以看出，将此类用户判成非离网是必然的。

(6) 应用模型。从前面的工作中可以得出一些简单的结论，比如通信支出越少的客户越容易流失、欠费频率越高的客户越容易流失等。除此之外，数据挖掘人员还应配合业务专家，根据数据挖掘分析寻找流失的原因，并找出潜在的规律，对未来的客户流失进行预测，指导业务行为。

为了提高预测准确度，使用ETHINK基于Saprk研发的拥有自主知识产权的数据挖掘平台，使预测流失准确率达到50%，查全率达到60%，使预测周期提前到一个半月。

深度学习特征处理：特征处理是影响模型效果比较重要的因素，鉴于此，引进了深度学习算法，自动提取、

处理原始数据特征，极大丰富了模型输入。随机森林多分类器是一个包含多个决策树的分类器，算法效果优于决策树，但其算法实现难度比较高，商业数据挖掘工具（IBM SPSS Modeler）也未集成该算法，因此针对该问题，基于 ETHINK 自主实现了随机森林算法。

参 考 文 献

[1] 李文，程华良，彭耀，等. 基于 Spark 可视化大数据挖掘平台.[M]//系统仿真技术及其应用学术（第 15 卷），合肥：中国科学技术大学出版社, 2014.

[2] 汤小文，蔡庆生. 数据挖掘在电信业中的应用[J]. 计算机工程, 2004.

[3] 郑冰. 数据挖掘在电信客户关系管理中的应用[J]. 中国数据通信, 2009.

[4] 李宁. 数据挖掘在电信 CRM 中的应用研究[D]. 重庆：重庆大学, 2005.