大数据驱动的投诉预测模型

周文杰[1] 杨 璐[1,2] 严建峰[1,2]
(苏州大学计算机科学与技术学院 苏州 215006)[1] (香港城市大学创意媒体学院 香港 999077)[2]

摘 要 随着电信行业市场竞争的不断加剧,用户对服务质量要求逐步提高,导致用户投诉率不断攀升。在此情况下,通过准确预测用户投诉行为来降低用户投诉率成为运营商关注的重点。目前传统的投诉预测模型仅从分类算法和人工调研特征来讨论,而没有充分利用运营商的大数据。因此,提出了在Hadoop/Spark大数据平台上使用并行随机森林来构建用户预测投诉模型,它不仅用到了业务支持系统数据,而且还用到了运营支持系统数据和客服工单数据,并在此基础上进一步增加了反映用户相互关系的图特征和二阶特征。基于上海市某运营商数据的实验结果表明,利用多来源、高维度的特征来训练用户投诉预测模型的精度会明显高于传统方法,在此基础上有针对性地对目标用户采取安抚措施,可以降低用户投诉率,获得较高的商业价值。

关键词 大数据,投诉预测模型,特征工程,二阶特征,图特征,随机森林
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.039

Big Data-driven Complaint Prediction Model

ZHOU Wen-jie[1] YANG Lu[1,2] YAN Jian-feng[1,2]
(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)[1]
(School of Creative Media, City University of Hong Kong, Hong Kong 999077, China)[2]

Abstract Because of fierce competition in telecommunication (telco) industry, it is crucial to reduce customer complaint rate and improve customer services to improve competitive advantages for telecommunication companies. Thus, accurately predicting the complaint behaviors to reduce the complaint rates becomes one of the most important tasks for telco operators. Traditional complaint prediction models only focus on classification algorithms and artificial feature selection and do not release the full power of telco big data. In this paper, we proposed a big data-driven complaint prediction model on the Hadoop/Spark platform using efficient parallel random forests. To better explore the performance of the proposed method, we performed feature engineering not only on all data from business support system (BSS) and operations support system (OSS), but also on those from the customer service records (CSR). Moreover, several useful graph-based features and second-order features between the relationship of users were designed and used to enhance the predictive performance. Experiment results based on the practical data of the telco operator in Shanghai show that using more data sources and high dimension data to train the complaint prediction models make the prediction accuracy higher than the state-of-the-art algorithms. Based on the result, we took comfort measures on target users, which can make the lo-wer complaint rate of users and bring significant business value to the operator.

Keywords Big data, Complaint prediction model, Feature engineering, Second-order feature, Graph-based features, Random forest

1 引言

随着电信业重组和互联网即时通信软件的崛起,用户面临的电信服务选择越来越多,对运营商服务质量的要求也越来越高,运营商的用户投诉量也随之呈现猛增趋势。以上海某运营商为例,其全网活跃用户大概为453万,仅2014年4月份的用户投诉量高达2.9万左右,全年投诉量占全网人数的7.6%以上。用户投诉量的增加不仅增加了处理投诉的成本,而且使用户对运营商的满意度下降,可能导致用户提前离网,使运营商营收下降。因此,如何减少用户的投诉量及提高用户的满意度成为各大运营商关注的重点。

为了减少用户的投诉量,一方面可以从运营商自身出发,提高电信产品服务质量;另一方面可以从预测用户投诉行为入手,通过数据挖掘模型事先预测未来可能会投诉的用户名

到稿日期:2015-05-23 返修日期:2015-08-12 本文受国家自然科学基金(61373092,61033013,61272449,61202029),江苏省教育厅重大项目(12KJA520004),江苏省科技支撑计划重点项目(BE2014005)资助。
周文杰(1989-),男,硕士生,主要研究方向为大数据挖掘、推荐系统和机器学习;**杨 璐**(1982-),女,博士,副教授,CCF会员,主要研究方向为软件可靠性和机器学习;**严建峰**(1978-),男,博士,副教授,CCF会员,主要研究方向为并行计算和机器学习,E-mail:yanjf@suda.edu.cn(通信作者)。

单，分析其投诉原因，针对这些用户提前制定不同的营销方案。目前针对运营商用户投诉模型已有一些研究[1-4]，其中文献[1-3]主要是通过改进算法本身来提高预测模型的精度，例如赵业祯等人[1]提出了基于 Bb 信令的通用分组无线服务业务潜在投诉的预测方法，从客户信令分析入手，采用决策树(Decision Tree，DT)[5]算法作为分类器来预测用户投诉倾向。栾媛媛[2]、龙雯雯[3]等人相继提出了基于改进 BP 神经网络(Back-Propagation network，BP)[6]的用户投诉预测模型。这些方法主要从业务支持系统(Business Support System，BSS)[7]数据(BSS 数据)中抽取特征。另一些解决方案则通过获取影响投诉的特征来提升模型预测精度，例如魏红明等人[4]利用实地调研的方式获取有用特征输入模型，从提高特征信息的角度提升预测精度。这些研究主要是从分类算法和人工调研特征方面讨论用户投诉预测模型，在处理传统小规模数据时较为有效。然而运营商大数据具有的巨大的数据量和特征量使得这些方案不再适用。首先，这些分类算法处理速度较慢，无法有效实现大规模并行。其次运营商大数据包含信息复杂，潜在高阶特征众多，通过人工实地调研的方式费时费力。

针对上述方案的不足，本文首先提出基于业界广泛采用的 Hadoop[8]/Spark[11]技术搭建一个运营商大数据平台，在平台上进一步提出了大数据驱动的用户投诉预测模型。对比之前的投诉预测研究工作，本文提出的大数据驱动的模型有以下优点：从大数据角度，首先，本文使用的数据来源比传统预测模型更丰富，训练数据量更大。例如，利用大数据平台整合了业务支持系统(BSS)数据、运营支持系统[9](Operation Support System，OSS)数据(OSS 数据)和客服工单(Customer Service Records，CSR)数据(CSR 数据)。模型训练所使用的数据量远远大于之前研究工作中使用的部分信令数据或者部分 BSS 数据。其次，使用的特征维度更加丰富，如增加了图特征和二阶特征。最后，使用的训练数据提前量更短，实验证明它能有效提升用户预测模型的精度。

从系统角度，本文设计了一套从数据采集、特征抽取最后到模型训练的分布式框架系统，其可以有效地处理运营商大数据，具有较好的推广性。在大数据平台上采用随机森林(Random Forest，RF)[12]作为分类器构建用户投诉预测模型。实验结果表明，提出的用户投诉预测模型在更大训练数据量、更高维度、更短提前量的数据情况下预测精度更高，运算速度更快。

本文将详细介绍大数据驱动下的用户投诉预测模型，第 2 节介绍大数据驱动下用户投诉预测模型的框架结构；第 3 节阐述用户投诉预测模型的特征选择；第 4 节介绍模型预测算法的设计；第 5 节给出实验结果，并对不同算法、不同特征下的实验结果加以分析；最后总结全文，并指出进一步的研究方向。

2 投诉预测大数据平台架构

本文选用了 Spark 分布式处理框架，其采用基于内存的分布式计算系统，建立在统一抽象的弹性分布式数据集[14](Resilient Distributed Datasets，RDD)之上，使其以基本一致的方式应对不同的大数据处理场景，其性能最高达到 Hadoop 的 100 倍左右。本文使用了上海市某运营商的真实用户数据，利用这些数据在一个多节点的 Hadoop/Spark 分布式平台上搭建了提出的用户投诉预测模型。该模型的框架如图 1 所示，最底层为数据源层，它为模型提供最原始的用户文本数据；第 2 层为数据存储层，它将数据源层的文本数据按天分区存储在 Hadoop 分布式文件系统[19](Hadoop Distributed File System，HDFS)上，并按不同类型存储在不同 Hive[17]表中；第 3 层为数据特征层，它将从数据存储层中生成原始特征、图特征和二阶特征；最上层为数据应用层，它将数据特征层中得到的特征数据输入分类器中来生成预测用户投诉名单和特征排名，并将特征排名反馈给数据特征层以进行特征的修改和删除。

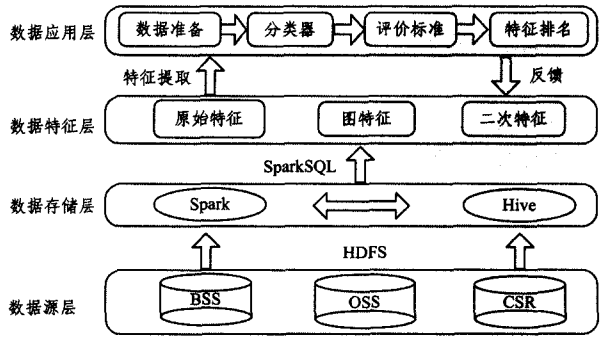

图 1 投诉预测大数据平台框架

2.1 数据源层

数据源层为用户投诉预测模型提供最原始的用户数据，其中包括 BSS 数据、OSS 数据和 CSR 数据 3 类数据。

1)运营商 BSS 数据：该组数据主要包含用户的一些基本信息、每星期用户的消费账单信息和消费行为信息，例如：用户的身份基本信息、用户每星期的余额信息等。

2)运营商 OSS 数据：该组数据主要包含运营商给用户提供的网络服务层信息。它大致分为电路交换(Circuit Switch，CS)数据、分组交换(Packet Switch，PS)数据。电路交换数据主要描述电话连接质量相关的信息，例如：用户掉话率和接通率等。分组交换数据主要描述通过深度报文检测(Deep Packet Inspection，DPI)解析出来的分类网络信息，例如：上网速度、联网成功率、HTTP 信息、WAP 信息等。

3)运营商 CSR 数据：该组数据主要包含运营商客服系统在受理用户请求和整个处理过程的数据，其中包含工单基本信息数据、工单环节信息数据，例如：过去几个星期用户投诉类工单数量、工单紧急程度、工单满意程度等。

2.2 数据存储层

数据存储层为原始用户数据提供了存储空间。这一层采用了一个多节点的 Hadoop/Spark 平台，每个节点都挂载了 12 个 2TB 的硬盘。首先将用户原始数据按天分区上传到多节点的 HDFS 上，其次根据 HDFS 上不同的类型数据创建对应 Hive 表，并且为每张 Hive 表添加按天分区字段，以加快查询速度。它为上层的数据特征层提供了数据查询支持。

2.3 数据特征层

数据特征层为投诉预测模型提供特征输入，它利用 SparkSQL[15]和现有算法对 BSS 数据、OSS 数据和 CSR 数据进行分析，生成了 3 大类特征。1)原始特征：这类特征是利用 SparkSQL 对源数据进行简单、复杂和统计查询得到的特征，

其中包括 BSS 特征、OSS 特征和 CSR 特征。2)图特征：这类特征是利用 Page Rank[22] 和 Label Propagation[23] 算法通过用户的通话量和短信量生成用户影响力特征和 Lp 特征。3)二阶特征：这类特征是利用因式分解机[24](Factorization Machines，FM)算法来找出潜在重要的特征对。

2.4 数据应用层

数据应用层是整个模型的最终实现层。它首先对数据特征层提取出来的特征样例做数据标注，将当前星期投诉过的用户标为正样例，将当前其余活跃用户标为负样例输入模型。随后将准备好的训练样例输入到随机森林分类器中进行训练，得到模型评估指标和特征排名。最后将特征排名反馈给数据特征层进行特征的修改和删除，进行下次模型训练。按照该模式，反复地优化用户投诉模型，直到达到预定精度目标。在实际上线时，只要将要实际预测的用户特征输入到优化好的模型中，就能得到用户投诉的可能性排名。

3 特征工程

特征工程是指利用目标问题所在特定领域的知识或者自动化的方法来生成、提取、删减或者组合变化得到的特征的处理步骤。在大数据分析中，模型的数据越来越多，训练时间也越来越长。如何从大量复杂特征中去除不相关特征并提取关键特征显得必不可少。这样不但使得模型的运算时间变短，而且提高了模型的精度，因此在建立基于大数据的投诉预测模型时，特征工程显得尤为重要。

3.1 原始特征

该类特征主要包括利用 SparkSQL 对 BSS 数据、OSS 数据和 CSR 数据进行简单、复杂和统计查询得到的 BSS 特征、OSS 特征和 CSR 特征。

BSS 特征包含了用户基本信息、每星期用户的消费账单信息和消费行为信息(如用户收到骚扰电话和骚扰短信的详单)等。这些信息在运营商存储表中容易获取，也是传统投诉模型常用的特征。本文通过关键绩效指标(Key Performance Indicators，KPI)[20] 和关键质量指标(Key Quality Indicators，KQI)[21] 选择了相对比较重要的 28 个特征。

OSS 特征因 OSS 数据量巨大而时常被传统预测模型忽略。本文同样利用 KPI/KQI 从 OSS 数据中的电路交换数据和分组交换数据里挖掘一些 OSS 特征，并对其做了一些简单计算和统计。其中电路交换特征是用户对运营商通话质量的反映。通话质量是用户对运营商的基本要求，所以本文对用户语音中的平均掉话率、平均语音质量、平均信号强度等 8 个 KPI/KQI 的电路分组特征进行了统计。分组交换特征是用户对运营商网络服务质量的反映。随着移动用户数量的增长，用户对移动网络要求越来越高。因此从用户使用最多的 Http、Wap、Email 方面入手，统计了与这 3 个方面相关的统计特征，例如：平均网页响应成功率、平均网页延迟时间和 Web 停顿次数等，最终总共选择了 15 个 KPI/KQI 的分组交换特征。

CSR 特征是投诉预测模型中最重要的特征，它主要来源于运营商客服管理系统，其信息主要包括用户过往投诉的记录、用户提交工单的属性等，它在一定程度上反映了当前用户的需求状态。本文参考褚卫艳[10] 的研究，从投诉工单相关属性中选择了比较重要的 15 个特征。表 1 列出了从 BSS 特征、OSS 特征和 CSR 特征中选出的 10 个主要特征。

表 1 部分原始特征

BSS 特征	OSS 特征	CSR 特征
年龄	平均掉话率	工单紧急程度
GPRS 流量	平均上行速度	工单流转标志
语音时长	平均下行速度	工单满意程度
入网时长	GET 首次延迟和	工单分发次数
信用等级	停顿次数均值	工程处理时间
与移动通话时长	平均通话质量值	过去一个星期咨询工单数
漫游通话分钟数	平均信号强度值	过去一个星期故障申告数
拨打人工服务次数	平均网页延迟时间	过去一个星期业务工单数
过去一周骚扰短信数	TCP 连接建立时长均值	过去一个星期营销工单数
过去一周骚扰电话数	平均网页响应成功率	过去一个星期投诉工单数

3.2 图特征

该类特征主要描述用户之间相互影响的关系程度。假设全网当前星期活跃用户数量 N，则利用全网所有用户之间一个星期内通话和短信次数建立起用户关系无向图，即构造两个 $N\times N$ 的对称矩阵，一个表示通话量的矩阵 M_v，另一个表示短信量的矩阵 M_m。如图 2 所示，M_v 中第 2 行第 3 列的 23 表示在当前一个星期内用户 2 和用户 3 的通话次数为 23；在 M_m 中则表示用户 2 和用户 3 当前一星期内的短信联系次数。得到这两个矩阵后，本文可利用现有的 Page Rank 和 Label Propagation 算法来构造关于通话量和短信量的影响力特征和 Lp 特征。

$$M_v=\begin{bmatrix} 0 & 11 & 4 & 5 & 8 & \cdots & 9 \\ 11 & 0 & 23 & 7 & & & \vdots \\ 4 & 23 & 0 & & \ddots & & \vdots \\ 5 & 7 & & \ddots & & \ddots & \vdots \\ 8 & & \ddots & & \ddots & & \vdots \\ \vdots & & & \ddots & & \ddots & \vdots \\ \vdots & & & & \ddots & & \vdots \\ 9 & \cdots & \cdots & \cdots & & \ddots & 0 \end{bmatrix}_{N\times N}$$

图 2 通话矩阵

用户影响力特征主要描述的是用户在其朋友圈内的重要程度，用户影响力越高，说明该用户在其朋友圈中相对其他人的语音、短信和上网流量等业务量都高。本文首先将所有的用户影响力设为 1，随后每个用户的影响力根据 Page Rank 中的公式来迭代更新，直到所有用户的影响力值都近似收敛为止。影响力更新公式如下：

$$I_i=\frac{(1-d)}{N}+d\sum_{j\in Nei(i)}\frac{I_j w_{i,j}}{\sum_{j\in Nei(i)} w_{i,j}} \tag{1}$$

其中，d 为阻尼系数，默认为 0.85；N 为当前一星期全网活跃用户数；$Nei(i)$ 为用户 i 所有通话或短信的朋友集合；$w_{i,j}$ 为通话或短信矩阵中第 i 行第 j 列的值。最后可得到所有用户关于通话量和短信量的影响力值。

Lp 特征主要描述未投诉用户受其已投诉用户好友的影响程度，Lp 值越高说明该用户受其已投诉朋友的影响越大，该用户也投诉的概率就越高。首先生成一个 $N\times 2$ 的矩阵 $L_{N\times 2}$，每一行表示一个用户投诉和不投诉的概率，当前一星期投诉过的用户表示为[1,0]，没投诉的则表示为[0,1]。随后按照下面的更新方式来迭代：

1) $L_{N\times 2}\leftarrow M_{N\times N}L_{N\times 2}$；

2)$L_{N\times 2}$中投诉过的用户不变，将没投诉过的用户的行归一化成概率，其相加等于1；

3)重复执行以上两步，直到$L_{N\times 2}$收敛。

其中，$M_{N\times N}$是描述用户通话量或短信量的矩阵。最后通过$L_{N\times 2}$得到未投诉用户通过标签传播后会投诉的倾向概率。

通过上述描述可得到关于描述用户之间关系的4个重要图特征：通话影响力特征、语音影响力特征、通话Lp特征和语音Lp特征。

3.3 二阶特征

该类的二阶特征由两个特征值相乘得到，这是由于有时单个特征的重要性不高，而两个特征的组合特征重要性却很高，后文中表3和表4的实验结果都验证了这一假想。本文利用现有FM算法来找取重要的特征对，FM的目标函数如下：

$$y=w_0+\sum_{i=1}^{n}w_i x_i+\sum_{i=1}^{n}\sum_{j=i+1}^{n}\langle v_i,v_j\rangle x_i x_j \tag{2}$$

其中，n为模型的总特征数；w_0，w_i为特征权重；v_i，v_j为矩阵$V_{n\times k}$中第i行和第j行的$1\times k$维的行向量，即用v_i，v_j来描述第i，j个特征的潜在关系，用$\langle v_i,v_j\rangle$代表第i，j个特征组合的权重。利用随机梯度下降使得该目标函数收敛后得到矩阵$V_{n\times k}$。随后将$V_{n\times k}$中的每一行两两求内积得到$\frac{(n+1)\times n}{2}$个实数，然后将其从大到小排列，实数越大说明其对应组成的二阶特征越重要。最终将排名前25的实数对应的特征对作为新的二阶特征加入到模型中。

4 预测算法设计

本节首先简要阐述投诉预测模型的分类器算法，其次描述该分类器算法在模型架构中如何实现并行计算和得到用户投诉预测结果。

4.1 随机森林算法

随机森林算法是一种统计学习算法，它利用boostrap[13]重抽样方法从原始样本中抽取多个样本，对每个boostrap样本进行决策树建模，然后组合多棵决策树的结果，通过每棵决策树的预测值进行投票(当模型为分类时)或求平均值(当模型为回归时)来最终决定预测值。基于用户投诉模型，假设每个用户给出的训练样本特征为$X_u=\{x_1,\cdots,x_i\cdots,x_n\}$，标签为$y_u=\{$投诉$=1$，没投诉$=0\}$。随机森林会训练多棵决策树$f_k(1\leqslant k\leqslant K)$，每棵决策树的特征只选择特征库中的一部分。对于样本X_u最终预测的标签结果是对所有K棵决策树的结果求平均值：

$$y_u=\frac{1}{K}\sum_{k=1}^{K}f_k(X_u) \tag{3}$$

其中，y_u为用户u下星期会投诉的可能性；K为随机森林中决策树的棵数；对于随机森林中的每棵决策树而言，会采用boostrap方法从所有样本中有放回地选择与原始样本等量的训练样本，从所有n个特征中随机选择$\sqrt{n}$个特征来训练决策树模型。而分支规则采用基尼[18](Gini)增长系数来划分，并且整个随机森林最后还能通过基尼增长系数得到所有特征的重要性，为数据特征层提供反馈信息。

4.2 参数并行学习框架

本文采用了一个多节点的Hadoop/Spark计算平台。由于随机森林训练的决策树棵数能达到上百棵，为了加快训练速度，本文将决策树训练任务平均分配到多个节点同时进行，如图3所示。首先将训练数据和模型源代码从本地上传到HDFS平台。当开始训练模型时，这m个节点从HDFS上下载源码和训练数据，m个节点并行地执行模型训练。当节点i外的其他节点训练结束后，就将结果发送到事前设定好的节点i上，等所有节点的结果都发送完毕后，该节点i将所有结果汇总求平均值，得到最终的预测结果。由于m个节点的训练数据和决策树棵数都是等量的，因此节点i不会出现忙等待的情况，从而缩短了训练时间。

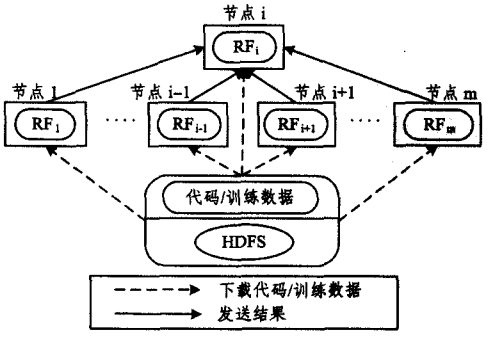

图3 并行计算框架

用户投诉预测模型的输入为描述用户的高维度特征向量，经过模型并行训练后输出每个用户对应下星期投诉的可能性。将其可能性从大到小排序，本文认为前U个用户为下星期会投诉的用户，即为正样例，其余则是不会投诉的用户，即为负样例。

5 实验结果与分析

本文实验基于多节点的Hadoop/Spark平台，每个节点的基本配置为：内存128GB，CPU Intel E5-2620 24 * 2.0 GHz。数据集为上海市某运营商2014年4月到2014年11月从BSS、OSS和CSR中抽取的真实数据，其中样本量为4532945，正负样例比例为1∶150。在特征上，使用上述3大类特征：原始特征、图特征和二阶特征，维度为95。每日原始数据量的大小如表2所列。

表2 每日数据量大小(GB)

数据源	数据类别	大小
BSS数据	用户行为	2.5
BSS数据	用户账单	0.8
BSS数据	语音/短信	18.3
OSS数据	CS数据	34
OSS数据	PS数据	1013
CSR数据	工单数据	2

5.1 评价标准

本文的评价指标是分类预测模型中常用的评价指标：召回率(Recall)、准确率(Precision)和AUC(Area Under Curve，AUC)。召回率是指预测用户投诉名单中投诉的人数与实际真实投诉人数的比率。准确率是指预测用户投诉名单中投诉的人数与预测用户投诉名单总人数的比率。AUC是指计算ROC(Receiver Operating Characteristic，ROC)[26]曲面下的面积，它能综合评价投诉预测模型的好坏。召回率的定义如下：

$$Recall=\frac{U_c}{C} \tag{4}$$

准确率的定义如下：

$$Precision=\frac{U_c}{N} \tag{5}$$

AUC 的定义如下：

$$AUC=\frac{\sum Rank-\frac{C\times(C+1)}{2}}{C\times NC} \tag{6}$$

其中，U_c 表示在排序预测名单中排名前 U 的用户中下星期真实投诉的人数；C 表示下星期真实投诉的人数；NC 表示下星期活跃用户中没有投诉的人数；$\sum Rank$ 表示将所有在网活跃用户经过模型得到的可能性从大到小排序，然后令最大可能性对应的用户排名为 N，第二大可能性对应的用户排名为 $N-1$，以此类推，将所有真实投诉用户的排名相加得到的值。

5.2 分类器精度评估

本文对比了几种分类算法：随机森林算法、逻辑回归[25](Logistic Regression，LR)算法、决策树算法和改进 BP 神经网络[2]算法。其中逻辑回归是目前工业界应用最广泛的分类器算法，而决策树和 BP 神经网络算法是传统用户投诉预测模型用到的分类器算法。

为了公平起见，这 4 个分类器算法的输入特征都是 3.1 节中的 BSS 特征和 OSS 特征，用当前一星期的用户数据训练预测下个星期可能投诉的用户名单，按照模型输出的用户投诉可能性从高到低排序，取前 10000 用户为正样例。对于随机森林参数，本文采用了 500 棵决策树，叶子节点最少样本为 100，这样可以避免过拟合。由于训练数据中正负样本不平衡，样本权重值设置为 150∶1；对于逻辑回归，由于它的输入更适合稀疏二进制的特征，因此在训练之前先对输入数据进行离散化处理，学习率为默认值 0.1；由于改进 BP 神经网络算法对初始网络权重非常敏感，因此本文经过反复调试得到了以下最优的结果。具体算法的实验效果和运行时间对比如图 4 所示，可以看出决策树在处理大规模连续值时预测效果不是很好，实际中它容易产生过拟合，但其运算时间相对较短。逻辑回归预测效果要好于决策树，但其在训练前要先做好离散化，过程比较繁琐，花费时间较长。改进 BP 神经网络在速度方面有了明显提高，但精度却远不如随机森林算法。实验结果表明在大数据情况下，相对其他 3 种算法，随机森林算法的实验效果最好。

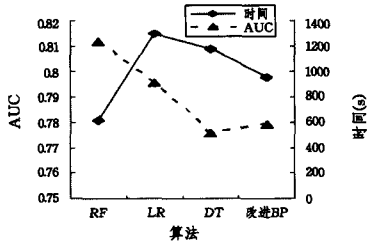

图 4 各种分类算法的 AUC 和时间对比

5.3 特征的评估

为了验证各类特征的重要性，本文在用户数据量不变的情况下，向仅使用 BSS 特征(B)的模型中不断加入新的特征。可以发现每当加入新的特征后，模型的召回率、准确率和 AUC 都会有不同程度的提升，加入 CSR 特征(CSR)和 OSS 特征(O)时提升最为明显，其次是图特征(GF)和二阶特征(SOF)。从表 3 可以看出，当加入 OSS 特征后，模型的 AUC 提升了 2.2%，准确度提升了 4.1%，说明传统模型没用到 OSS 特征，该特征对模型的准确度提升有一定的帮助；当加入一星期的 CSR 特征后，模型的 AUC 提升高达 2.5%，准确率提升8.5%，由此说明用户之前的投诉信息能很好地描述当前想要投诉用户的特征；当加入图特征后，模型 AUC 也有了 1.2%的提升，准确度有了 1.3%的提升；当加入二阶特征后，模型的 AUC 和精度也有了稍微的提升，这说明利用现有算法得到已有特征间隐藏关系的二阶特征也能提高模型的精度。

表 3 各种特征加入后的结果

特征属性	召回率	准确率	AUC
B	0.2809	0.3856	0.7956
B+O	0.2901	0.4012	0.8136
B+O+CSR	0.3178	0.4356	0.8329
B+O+CSR+GF	0.3209	0.4412	0.8432
B+O+CSR+GF+SOF	**0.3429**	**0.4497**	**0.8469**

表 4 是通过所有特征训练模型后输出的前 10 名的重要特征。其中 BSS、OSS 和 CSR 表示特征来源于原始特征，GF 表示特征来源于图特征，SOF 表示特征来源于二阶特征。可以看出最重要的是 CSR 特征，其次是图特征和 OSS 特征，这也为表 3 中 AUC 和准确率的提升提供了有力的依据。

表 4 重要性排名前十的特征

排名	特征	特征类别	重要性
1	过去 2 月投诉工单数量	CSR	0.0925
2	TCP 连接建立时长均值	OSS	0.0588
3	工单流转标志	CSR	0.0523
4	语音影响力	GF	0.0500
5	过去一月骚扰电话量	BSS	0.0490
6	TCP 建立时长均值十漫游呼出次数	SOF	0.0441
7	语音 Lp	GF	0.0287
8	工单处理时间	CSR	0.0248
9	年龄	BSS	0.0223
10	工单满意程度	CSR	0.0219

5.4 提前量对精度的影响

为了验证提前量对精度的影响，本文分别用一个月、两个星期和一个星期作为提前量来预测用户投诉。即用当月的用户数据来预测下个月用户投诉的精度，用前两星期的用户数据来预测后两星期的用户投诉精度和用当星期的用户数据来预测下星期用户投诉的精度。从图 5 可以看出提前量为一星期的效果会好于一个月的效果，说明更短的提前量数据对模型的精度有更大的帮助。经分析，这是由于对于用户投诉问题来说，它具有很强的时效性。

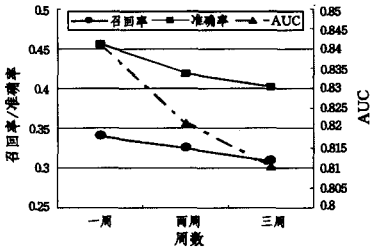

图 5 各提前量结果对比

5.5 数据量对精度的影响

为了验证数据量对精度的影响，本文分别用 1 个星期的用户数据量到 5 个星期的用户数据量来预测下个星期的用户投诉。图 6 的实验结果表明，随着训练数据量的增加，模型的精度会有一定的提升。但到了 3 个星期之后，可以发现模型的 3 个指标都变化不大，最后趋于一种稳定状态。

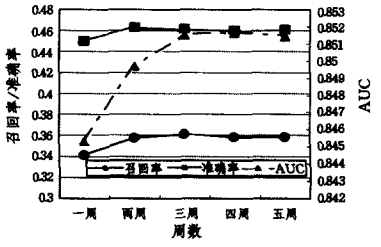

图 6　各种数据量结果对比

5.6　实际运行效果

本文提出的运营商用户投诉模型被实际运用到了上海市某运营商的大数据项目中，在实际上线运行中体现了良好的预测性能。在模型给出的前 5000 个预测用户的投诉名单中的准确率高达 91%。运营商的外呼部门通过模型给出的预测用户投诉名单，提前针对每个用户进行及时的安抚和关怀。运营商的技术部门也通过对模型给出的特征权重排名进行分析来优化电信产品服务。最终不但使得运营商客服外呼团队人员投入减少了 25%，而且使得该运营商每月的用户投诉率降低了 14.5%，用户离网率也随之下降了 5.3%，提升了用户的满意度。除此之外，基于用户投诉模型的大数据平台也为大数据项目的离网预测模型、电信产品推荐模型和基站投资模型提供了平台支持。

结束语　随着大数据在各行各业的广泛应用，本文在大数据的背景下运用大数据来预测上海市某运营商用户投诉行为。首先通过不断加入新的数据来源、新的特征和更短提前量的数据来证实利用数据量更大、数据维度更高和提前量更短的数据对用户投诉预测模型的性能有很大的提升。其次通过现有算法设计的图特征和二阶特征也能有效提升用户投诉预测模型的精度，降低用户投诉率，提高运营商的服务质量。

本文使用人工特征分析的方法来进行特征工程，能够根据应用的特点有效地抽取特征，在实践中获得了很好的应用效果。然而人工特征工程的方法需要有较多的行业经验积累和人力成本投入，有一定的应用限制。注意到了深度学习(Deep Learning，DP)[16]在自动获取有效特征方面的应用潜力，它能够自动学习要建模的数据的潜在分布的多层表达，提取分类需要的低层次和高层的特征，但在本项目中由于该模型的层次复杂、参数过多导致其学习速度太慢，最终本文没有使用该算法来进行特征提取。但随着深度学习被不断深入研究和优化，相信未来它在自动特征提取方面会发挥越来越重要的作用。

参 考 文 献

[1] Zhao Ye-zhen, Huang Xiao-di. Potential customer complaints predict model based on GPRS signaling[J]. Telecommunications Information: Network and Communication, 2014(8): 29-32(in Chinese)
赵业祯，黄晓弟. 基于信令的 GPRS 潜在投诉客户预测模型[J]. 电信快报：网络与通信，2014(8)：29-32

[2] Luan Yuan-yuan, Wang Zhong-ren, Xi A-dan, et al. Research on customer complaints warning model based on improved BP neural network[C]// Conference of Chinese Institute of Communications, 2010(in Chinese)
栾媛媛，王忠仁，奚阿丹，等. 基于改进 BP 神经网络的客户投诉预警模型研究[C]// 中国通信学会学术年会. 2010

[3] Long Wen-wen. Research on mobile user's complaint behavior based on Data Mining[D]. Chongqing: Chongqing University of Technology, 2014(in Chinese)
龙雯雯. 基于数据挖掘的移动用户投诉行为研究[D]. 重庆：重庆理工大学，2014

[4] Wei Hong-ming. Research on prediction model of data mining based on mobile communication customer complaints [D]. Hengyang: University of South China, 2009(in Chinese)
魏红明. 基于数据挖掘的移动通信客户投诉预测模型研究[D]. 衡阳：南华大学，2009

[5] Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1): 80-108

[6] W Jun-qing. BP Neural Network and Its Improvement[J]. Journal of Chongqing Institute of Technology(Natural Science Edition), 2007

[7] Shimada T, Akita K. Business support system[P]. US, US6868390 B1, 2000

[8] Yang Chen-tao. Data mining based on Hadoop[D]. Chongqing: University of Chongqing, 2010(in Chinese)
杨宸铸. 基于 HADOOP 的数据挖掘研究[D]. 重庆：重庆大学，2010

[9] Bhushan B, Hall J, Kurtansky P, et al. Operations Support System for End-to-End QoS Reporting and SLA Violation Monitoring in Mobile Services Environment[J]. Quality of Service in the Emerging Networking Panorama, 2004, 3266: 378-387

[10] Chu Wei-yan. The design of forecasting system based on the analysis of historical complaint data [D]. Beijing: Beijing University of Posts and Telecommunications, 2013(in Chinese)
褚卫艳. 基于投诉历史数据的分析和预测系统设计[D]. 北京：北京邮电大学，2013

[11] Luo Y, Wang W, Lin X. SPARK: A Keyword Search Engine on RelationalDatabases[C]// IEEE 24th International Conference on Data Engineering, 2008(ICDE 2008). 2008: 1552-1555

[12] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32

[13] MacKinnon D P, Williams C M L & J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods[J]. Multivariate Behavioral Research, 2004, 39(1): 99-128

[14] Zaharia M, Chowdhury M, Das T, et al. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing[C]// USENIX Symposium on Networked Systems Design and Implementation (NSDI). 2012: 141-146

[15] Gao Yan-jie, Chen Guan-cheng. SparkSQL: Big data processing engine based on memory[J]. Programmer, 2014(8): 104-107(in Chinese)
高彦杰，陈冠诚. SparkSQL：基于内存的大数据处理引擎[J]. 程序员，2014(8)：104-107

[16] Guo Lil-i, Ding Shi-fei. The research progress of Deep Learning [J]. Computer Science, 2015, 42(5): 28-33(in Chinese)
郭丽丽，丁世飞. 深度学习研究进展[J]. 计算机科学，2015，42(5)：28-33

[17] Thusoo A, Sarma J S, Jain N, et al. Hive-A Warehousing Solution Over a Map-Reduce Framework[C]// Proceedings of the Vldb Endowment(VLDB'09). 2009

[18] Dorfman R. A Formula for the Gini Coefficient[J]. Review of Economics and Statistics, 1979, 61(1): 146-149

[19] Mackey G, Sehrish S, Wang J. Improving metadata management

for small files in HDFS[C]// IEEE International Conference on Cluster Computing and Workshops, 2009 (CLUSTER'09). IEEE, 2009: 1-4

[20] Chambers D W. Key performance indicators[J]. Journal of the American Dental Association (JADA), 2013, 144(3): 242-244

[21] Simundic A. Quality indicators[J]. Biochemia Medica, 2008, 18 (3): 311-319

[22] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[J]. Stanford InfoLab, 1998, 9 (1): 1-14

[23] Kang F, Jin R, Sukthankar R. Correlated Label Propagation with Application to Multi-label Learning[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006: 1719-1726

[24] Rendle S. Factorization Machines[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). 2010: 995-1000

[25] Fan R, Chang K, Hsieh C, et al. LIBLINEAR: A Library for Large Linear Classification[J]. Journal of Machine Learning Research, 2008, 9(12): 1871-1874

[26] Mh Z. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine[J]. Clinical Chemistry, 1993, 39(4): 561-577

(上接第 196 页)

[5] Wang Hai-xun, Wang Wei, Yang Jiong, et al. Clustering by Pattern Similarity in Large Data Sets [C]// Proceedings of the 28th ACM SIGMOD International Conference on Management of Data. 2002: 394-405

[6] Pensa Ruggero G, Boulicaut Jean-Francois. Constrained Co-clustering of Gene Expression Data[C]// Proceedings of the 8th SIAM International Conference on Data Mining (SDM). 2008: 25-36

[7] Faris A, Bader Joel S, Rajul A, et al. Query-based Biclustering using Formal Concept Analysis [C]// Proceedings of the 12th SIAM International Conference on Data Mining (SDM). 2012: 648-659

[8] Jiang Tao, Li Zhan-huai, Chen Qun, et al. Towards OrderPreserving SubMatrix Search and Indexing [M]// Database Systems for Advanced Applications: Proceedings of the 20th International Conference on Database Systems for Advanced Applications (DASFAA) Part II, 2015: 309-326

[9] Jiang Tao, Li Zhan-huai, Shang Xue-qun, et al. Constrained Query of Order-Preserving SubMatrix in Gene Expression Data [J]. Frontiers of Computer Science, 2016, 10(5): 1-5

[10] Wassim A, Mourad E, Hao Jin-kao. BicFinder: a Biclustering Algorithm for Microarray Data Analysis [J]. Knowledge and Information Systems, 2012, 30(2): 341-358

[11] Yang Jiong, Wang Wei, Wang Hai-xun, et al. δ-Clusters: Capturing Subspace Correlation in a Large Data Set [C]// Proceedings of the 18th International Conference on Data Engineering (ICDE). IEEE press, 2002: 517-528

[12] Cho H, Dhillon Inderjit S, Guan Yu-qiang, et al. Minimum Sum-Squared Residue Co-clustering of Gene Expression Data [C]// Proceedings of the 4th SIAM International Conference on Data Mining (SDM). SIAM Press, 2004: 114-125

[13] Chen Shu-hua, Liu Juan, Zeng Tao. MMSE: A Generalized Coherence Measure for Identifying Linear Patterns[C]// Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE press, 2014: 489-492

[14] Matteo D, Alessandro F, Manuele B. Biclustering Gene Expressions using Factor Graphs and the Max-sum Algorithm[C]// Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI). AAAI Press, 2015: 925-931

[15] Zhao Yu-hai, Wang Guo-ren, Yin Ying, et al. A Novel Approach to Revealing Positive and Negative Co-Regulated Genes [J]. Journal of Computer Science and Technology, 2007, 22(2): 261-272

[16] Chen Jiun-rung, Chang Ye-in. An of Up-Down Bit Pattern Approach to Coregulated and Negative-Coregulated Gene Clutering of Microarray Data[J]. Journal of Computational Biology, 2011, 18(12): 1777-1791

[17] Zhao Yu-hai, Yu Xu, Wang Guo-ren, et al. Maximal Subspace Coregulated Gene Clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(1): 83-98

[18] Wang Guo-ren, Yin Lin-jun, Zhao Yu-hai, et al. Efficiently Mining Time-Delayed Gene Expression Patterns [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2010, 40(2): 400-411

[19] Wang Guo-ren, Zhao Yu-hai, Zhao Xiang-guo, et al. Efficiently Mining Local Conserved Clusters from Gene Expression Data [J]. Neurocomputing, 2010, 73(7-9): 1425-1437

[20] Yin Ying, Zhao Yu-hai, Zhang Bin, et al. Mining Synchronous and Asynchronous Co-Regulated Gene Clusters from Time Series Microarray Data [J]. Chinese Journal of Computers, 2007, 30(8): 1302-1314(in Chinese)
印莹, 赵宇海, 张斌, 等. 时序微阵列数据中的同步和异步共调控基因聚类[J]. 计算机学报, 2007, 30(8): 1302-1314

[21] Amichai P, Saharon R. Optimal Set Cover Formulation for Exclusive Row Biclustering of Gene Expression [J]. Journal of Computer Science and Technology, 2014, 29(3): 423-435

[22] Rui H, Maderia Sara C. BicSPAM: Flexible Biclustering using Sequential Patterns [J]. BMC Bioinformatics, 2014, 15(1): 1-20

[23] Trapp Andrew C, Li Chao, Patrick F. Recovering All Generalized Order-Preserving SubMatrices: New Exact Formulations and Algorithms [J/OL]. Annals of Operations Research, 2016. http://link. springer. com/article/10. 1007% 2Fs10479-016-2173-9

[24] Jiang Tao, Li Zhan-huai, Chen Qun, et al. Parallel Partitioning and Mining Gene Expression Data with Butterfly Network [M]// Database and Expert Systems Applications: Proceedings of the 24th International Conference on Database and Expert Systems Applications (DEXA), Part I. 2013: 129-144

[25] Jiang Tao, Li Zhan-huai, Chen Qun, et al. OMEGA: An Order-Preserving SubMatrix Mining, Indexing and Search Tool [M]// Machine Learning and Knowledge Discovery in Database: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/ PKDD), Part III. 2015: 303-307

[26] Broad Institute. Datasets. rar and 5q_gct[DB/OL]. http://www. broadinstitute. org/cgi-bin/cancer/datasets. cgi

[27] Gao B J, Griffith O L, Ester M, et al. On the Deep Order-preserving Submatrix Problem: a Best Effort Approach [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 309-325