

# 基于深度学习的用户投诉预测模型研究<sup>\*</sup>

周文杰<sup>1</sup>, 严建峰<sup>1,2</sup>, 杨璐<sup>1</sup>

(1. 苏州大学 计算机科学与技术学院, 江苏 苏州 215006; 2. 香港城市大学 创意媒体学院, 香港 999077)

**摘要:** 用户投诉预测模型能有效地降低电信用户投诉率, 对企业提高用户满意度和竞争力有着至关重要的作用。在模型训练过程中, 由于人工设计特征的缺陷和设计过程中存在难以预估的复杂性, 使得模型预测的精度和设计特征的效率不能有很大的提升。针对上述问题, 提出了一种基于深度学习的用户投诉预测模型。该模型通过深层网络特征学习单元能从电信用户原始数据中自动学习到适合分类器分类的高层非线性组合特征, 并将这些高层特征输入到传统分类器中来提高模型的精度。通过实验结果分析, 预测模型在 AUC 指标上比以往用户投诉模型提升了 7.1%, 证明了该模型自动学习特征的有效性和深度学习在电信大数据领域的可用性。

**关键词:** 电信投诉预测; 深度学习; 深度置信网络; 非线性组合特征

**中图分类号:** TP181

**文献标志码:** A

**文章编号:** 1001-3695(2017)05-1428-05

doi:10.3969/j.issn.1001-3695.2017.05.034

## Research on prediction model of complaint based on deep learning

Zhou Wenjie<sup>1</sup>, Yan Jianfeng<sup>1,2</sup>, Yang Lu<sup>1</sup>

(1. School of Computer Science and Technology, Soochow University, Suzhou Jiangsu 215006, China; 2. School of Creative Media, City University of Hong Kong, Hong Kong 999077, China)

**Abstract:** User complaints prediction model could effectively lower complaints rate of telecom users which was crucial for enterprise to improve customer satisfaction and competitiveness. During the process of model training, due to the defects of artificial design features and the unpredictable complexity in the design process, the accuracy of model prediction and the efficiency of the design features couldn't have a lot of ascension. For the above problems, this paper proposes a user complaints prediction model based on deep learning. By deep network characteristics study session, the model could automatically study the senior nonlinear combination features from the original telecommunication users' data which was suitable for classifier classification and import high-level features into traditional classifier to improve the precision of the model. By analyzing the experimental results, in this paper, the prediction model's AUC index has increased by 7.1% compared to previous user complaints model. It has demonstrated that the effectiveness of the model automatic study features and availability of deep learning in telecom big data domain.

**Key words:** telecom complaints prediction; deep learning; deep belief nets; nonlinear combined features

## 0 引言

随着移动互联网技术的快速发展,用户对运营商产品服务的要求及维权意识也在不断提高。当产品质量出现问题时,用户就会通过各种途径进行相关问题的投诉。用户投诉量的增加,不仅增加了企业处理投诉的成本,而且使用户对该运营商的满意程度下降,可能导致用户提前离网转用其他服务,造成运营商营收下降。因此,如何减少用户的投诉量,提高用户的满意程度成为各大运营商关注的焦点。

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单,分析其投诉原因。在此基础上针对这些用户提前制定不同的营销方案,主动对该用户群进行及时的安抚和关怀,将投诉处理从原来的“事后弥补”变为“事前控制”,使得运营商在处理

投诉问题时由被动变为主动,从而大幅度地降低用户投诉率,增强运营商自身的竞争力。

目前针对运营商用户投诉预测模型已有一些研究,他们主要从特征选择和优化分类器模型两个方面来研究。特征选择方面,2009年魏红明<sup>[1]</sup>利用研究邵阳移动的数据通过实地调研和访谈得到影响用户投诉的动态因素和静态因素;通过数据挖掘方法的理论,应用 OLAP(on-line analytical processing)分析方法提取影响客户投诉的因素作为模型特征;最后利用上述得到的特征加入 BP 神经网络算法(back-propagation neural networks, BP)中进行移动用户投诉预测。2014年赵业祯等人<sup>[2]</sup>提出了基于 Gb 信令的 GPRS(general packet radio service)业务的潜在投诉用户预测方法。他们主要是从用户投诉的信令分析入手,归纳信令特征,建立信令特征库,最后采用决策树(decision tree, DT)算法作为分类器来预测用户投诉行为。优化分类器模型方面,2010年栾媛媛等人<sup>[3]</sup>提出了基于改进 BP 神经

**收稿日期:** 2016-04-11; **修回日期:** 2016-05-16 **基金项目:** 国家自然科学基金资助项目(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)

**作者简介:** 周文杰(1989-), 男, 江苏泰兴人, 硕士研究生, 主要研究方向为大数据挖掘、推荐系统和机器学习(20134227015@suda.edu.cn); 严建峰(1978-), 男, 江苏昆山人, 副教授, 博士, 主要研究方向为并行计算、深度学习; 杨璐(1982-), 女, 副教授, 博士, 主要研究方向为软件可靠性、机器学习。

网络的客户投诉预警模型。该文主要采用 LM (Levenberg-Marquardt, LM) 优化的学习方法使 BP 算法在很大程度上减少迭代次数,加快预测模型的速度。除了以上两方面,2015 年周文杰等人<sup>[4]</sup>提出了基于大数据驱动的投诉预测模型,他们首先在以往预测模型的数据集中加入大量运营支持系统 (operation support system, OSS)<sup>[5]</sup> 数据,运用现有机器学习算法自动学习能描述用户间关系的图特征和特征间的二阶组合特征,通过此类特征来提高模型的预测精度,最后运用并行随机森林 (parallel random forest, PRF) 来加快模型运行速度。

上述的方法都未能自动捕捉到一些非线性的组合特征来提高模型的精度。故本文创新地提出了基于深度学习 (deep learning, DL)<sup>[6]</sup> 的用户投诉预测模型,运用深层网络模型深度置信网络 (deep belief nets, DBN)<sup>[7]</sup> 来自动学习用户特征间的非线性组合特征。并通过实验结果验证了该模型自动学习特征的有效性。

## 1 用户投诉特征选择

本文所使用的数据主要包括业务支持系统 (business support system, BSS)<sup>[8]</sup> 数据、运营支持系统 (OSS) 数据和客服工单数据 (customer service records, CSR)。BSS 数据主要包含了用户的一些基本信息、每天用户的消费账单信息和消费行为信息。OSS 数据主要包含运营商给用户提供的网络服务层信息。CSR 数据主要包含运营商客服系统受理用户请求,并完成用户诉求处理整个过程的数据。

### 1.1 BSS 特征

BSS 特征包含了用户基本信息、每天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。在本文中,通过关键绩效指标 (key performance indicators, KPI)<sup>[9]</sup> 和关键质量指标 (key quality indicators, KQI)<sup>[10]</sup> 选择了相对比较重要的 40 个特征。

### 1.2 OSS 特征

OSS 数据在三大数据源中占的比重最大,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,并作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应。通话质量是用户对运营商的基本要求,所以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 15 个 KPI/KQI 的电路交换特征进行了统计。分组交换特征是用户对运营商网络服务质量的反应。随着移动用户的增长,用户对移动网络要求越来越高。所以本文从用户使用最多的 Http、Wap、Email 方面入手,统计了与这三个相关的统计特征。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 20 个 KPI/KQI 的分组交换特征。

### 1.3 CSR 特征

CSR 特征是投诉预测模型中比较重要的特征,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,它在一定程度上反应了当前用户的需求状态。从投诉工单相关属性中选择了其中比较重要的 30 个特征。表 1 中列出从 BSS 特征、OSS 特征和 CSR 特征中选出的 20 个主要特征。

## 2 基于深度学习的用户投诉模型

本章将介绍深度学习的深度网络结构模型:深度置信网络 (DBN) 的基本原理和训练步骤,详细介绍本文基于 DBN 模型的用户投诉预测模型框架流程。

表 1 部分用户特征

| BSS 特征          | OSS 特征       | CSR 特征 |
|-----------------|--------------|--------|
| 年龄              | 平均掉话率        | 转派次数   |
| 信用等级            | 上行流量均值       | 投诉现象   |
| 在网时长            | 下行流量和        | 投诉原因   |
| 是否上海户籍          | 下行流量均值       | vdn 编号 |
| 性别              | 上行包大小和       | 受理员工号  |
| ARPU 值          | 上行包大小均值      | 受理方式   |
| GPRS 流量         | 下行行包大小和      | 故障地点   |
| GPRS 费用         | 下行包大小均值      | 影响度    |
| 本地通话分钟数         | TCP 连接建立时长均值 | 优先级    |
| 长途通话分钟数         | GET 首次延迟和    | 服务级别   |
| 漫游通话分钟数         | GET 首次延迟均值   | 紧急程度   |
| 语音通话时长          | 页面延迟和        | 满意度    |
| 点对点短信上行条数       | 页面延迟均值       | 责任原因   |
| 点对点短信上行费用       | 页面延迟流量和      | 责任部门   |
| 每天骚扰电话数量        | 页面延迟流量均值     | 解决程度   |
| 每天骚扰短信数量        | 页面大小和        | 再处理次数  |
| 免费时长            | 页面大小均值       | 反馈次数   |
| 拨打 10010 次数     | 页面 sr 延迟均值   | 等待时长   |
| 拨打 10010 人工服务次数 | 平均信号强度值      | 延时申请状态 |

### 2.1 DBN 模型

受限玻尔兹曼机 (restricted Boltzmann machines, RBM) 是第一个用非监督学习算法来训练浅层网络每一层的模型。RBM 是玻尔兹曼机 (Boltzmann machines, BM) 的一种特殊拓扑结构。BM 的原理起源于统计物理学,它是一种基于能量函数的建模方法,能够描述变量之间的高阶相互作用关系。如图 1 所示, BM 是一种对称耦合的随机反馈型二值神经网络,顶层表示隐层  $h$  的单元特征向量,底层为可视层  $v$  的单元向量,链接权重  $w$  为模型参数。BM 的可视层与隐层之间、隐层与隐层之间、可视层和可视层之间都相互连接,通过权值表达单元之间的相关性。而 RBM 限制了 BM 同一层之间的连接,它是一类具有对称性,无自反馈的双层、无向随机神经网络模型。在给定的可视层  $v$  单元状态时,各隐层  $h$  单元之间条件独立;同样在给定的隐层  $h$  单元状态时,各可视层  $v$  单元之间条件独立。对于一组给定的状态  $(v, h)$ , RBM 所具备的能量计算式如下:

$$E(v, h) = - \sum_{i=1}^n a_i v_i - \sum_{j=1}^m b_j h_j - \sum_{i=1}^n \sum_{j=1}^m v_i w_{ij} h_j \quad (1)$$

其中:  $w_{ij}$  表示可视层  $i$  与隐层  $j$  之间的连接权重,  $a_i$  表示可视层  $i$  的偏置,  $b_j$  表示隐层  $j$  的偏置。设 RBM 的参数为  $\theta$ ,  $\theta = \{w_{ij}, a_i, b_j\}$ , 当参数  $\theta$  固定时,基于该能量函数即可得到  $(v, h)$  的联合概率分布为

$$P(v, h) = \frac{e^{-E(v, h)}}{Z(\theta)} \quad (2)$$

$$Z(\theta) = \sum_v \sum_h e^{-E(v, h)} \quad (3)$$

其中:  $Z(\theta)$  为归一化因子,对于实际问题需要求解联合概率分

布  $P(v, h)$  的边缘分布:  $P(v) = \frac{1}{Z(\theta)} \sum_h e^{-E(v, h)}$ , 由于 RBM 隐层和可视层之间存在条件独立和对称性。当给定  $v$  时,  $h_j$  的激活概率为:  $P(h_j = 1) = \sigma(b_j + \sum_i v_i w_{ij})$ , 当给定  $h$  时,  $v_i$  的激活概率为:  $P(v_i = 1) = \sigma(a_i + \sum_j h_j w_{ij})$ , 其中  $\sigma(z) = 1/(1 + \exp(-z))$ 。由此模型参数  $\theta$  可以通过最大化 RBM 在训练集上的对数似然函数得到模型的最优参数  $\theta^*$ 。

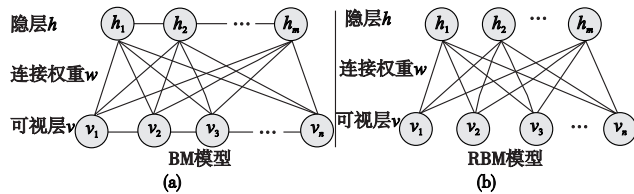


图1 BM与RBM模型结构图

深度置信网络 (DBN) 是以 RBM 为单元串联构建的深度学习方法, 训练时使用非监督贪心逐层算法进行。它也是个能够捕捉训练数据在隐层表现出来的高阶数据的概率生成模型。可视层  $v$  的单元特征向量  $x$  和隐层单元向量  $h^k$  (其中隐层为  $l$  层) 的联合分布为

$$P(x, h^1, h^2, \dots, h^l) = \left( \prod_{k=0}^{l-2} P(h^{k+1} | h^k) \right) P(h^{l-1}, h^l) \quad (4)$$

其中:  $x = h^0$ ,  $P(h^k | h^{k+1})$  表示第  $k+1$  个 RBM 隐层的可视层的条件概率分布,  $P(h^{l-1}, h^l)$  表示最顶层 RBM 中可视层和隐层的联合分布, 具体结构如图 2 所示。

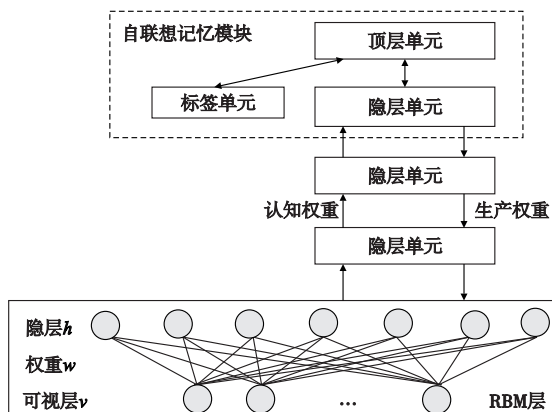


图2 DBN模型结构

DBN 的训练步骤如下:

- 把  $x = h^0$  作为可视层的输入训练第一个 RBM 模型。
- 低层的 RBM 使用观察值样本来进行训练, 固定好 RBM 的模型参数。
- 将低层 RBM 的输出作为高层 RBM 的输入进行训练。
- 重复执行 b) c), 重复训练所有的 RBM, 实现模型参数的初始化。
- 在确定了模型参数后, 在 DBN 自联想记忆模块利用带标签数据用 BP 算法对判别性能作微调, 这样通过一个自下而上反馈学习方法来调整所有 RBM 的模型参数。

在 DBN 所有模型参数都学习固定好的情况下, 向可视层输入原始特征 (BSS 特征、OSS 特征和 CSR 特征), 经过多隐层的自动学习获得更能表示数据特性的高层特征。这样就可以通过深度学习多层网络自动学习到有效高层特征。

## 2.2 基于 DBN 模型的用户投诉预测模型框架

近些年来深度学习被广泛应用于图像处理、语音识别、搜索引擎等许多领域, 它能够从大量数据中学习分类所需的高

层、抽象的特征表示。为了获得更多复杂非线性的组合特征, 本文将创新性地深度学习方法应用到电信大数据投诉预测问题中, 如图 3 所示为本文运用深度学习中 DBN 模型来研究投诉问题的流程框架。

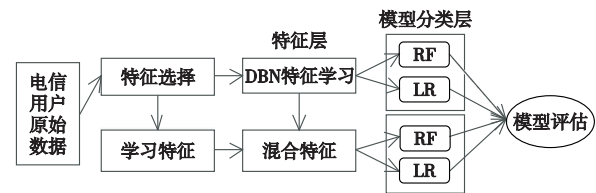


图3 深度学习研究框架

1) 从电信用户原始数据库中根据投诉相关的业务知识进行特征选择, 构建出模型的原始特征 (BSS 特征、OSS 特征和 CSR 特征)。然后将原始特征输入 DBN 特征学习模型, 进行多次的模型训练选择出最优的初始化参数设置, 在最优参数的模型训练下获得最高层的特征表达, 将其输入随机森林 (random forest, RF)<sup>[11]</sup> 和逻辑回归 (logistic regression, LR)<sup>[12]</sup> 分类器中进行模型的评估。

2) 将文献[4]中的学习特征与基于 DBN 模型自动学习的高层非线性组合特征混合进行投诉预测, 并与其他两种模型进行分析对比, 证实了 DBN 自动学习特征的有效性及其深度学习模型在电信大数据领域的可用性。

## 3 实验结果与分析

### 3.1 数据集与实验环境

本文使用的数据是上海市某运营商 2014 年 6 月到 2014 年 12 月从 BSS 数据库、OSS 数据库和 CSR 数据库中抽取的真实数据, 由于实际在网活跃用户中投诉用户相对不投诉用户还是占少部分, 这造成了正负样本不平衡。本文的负样本首先通过一些特定规则获得一些高价值无投诉用户, 再从其中随机抽取部分负样例, 最终挑选得到 1 032 945 个样本, 正负样本比例为 1:3。由于训练 DBN 耗时较长, 因此本文使用的是 Python 中的 Theano 库来训练 DBN。Theano 是根据 DBN 训练特点而创建的 Python 库, 它不仅操作简单, 而且可以通过调用图形处理器 (graphic processing unit, GPU) 加速整个训练过程, 本文用于训练 DBN 的步骤参考文献[13]。在实际训练过程中本文采用的硬件环境为: 内存 32 G, CPU Intel Core i7-4790 8 \* 3.60 GHz, GPU GeForce GTX Titan X。它的训练速度是使用单核 Intel 3.47GHz 的 15 倍以上。本文基于 DBN 的用户预测模型主要包含三个部分: a) 特征提取。本文采用的是第 2 章所描述的原始特征 (BSS 特征、OSS 特征和 CSR 特征), 总计 105 个。b) 特征预处理。由于用户原始特征维度跨度较大, 故本文将原始数据归一化到 [0, 1]。计算公式为

$$z_f = \frac{x_f - \min(f)}{\max(f) - \min(f)} \quad (5)$$

其中:  $\min(f)$ 、 $\max(f)$  表示特征  $f$  的最小值和最大值;  $x_f$  表示用户  $x$  的特征  $f$  的值;  $z_f$  表示用户  $x$  的特征  $f$  归一化后的值。DBN 特征学习: 将预处理好的特征直接输入到 DBN 模型中, 通过 DBN 模型自动学习高层非线性组合特征。最后将 DBN 学习好的特征输入到传统分类器。本文采用了随机森林 (RF) 和逻辑回归 (LR) 两种分类器, 采用曲线下面积 (area under curve, AUC)<sup>[14]</sup> 指标来衡量模型的预测效果, 具体 AUC 计算式为



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

其中: $C$ 表示即将真实投诉的人数; $NC$ 表示后续时间里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设为 $N$ ,第二大可能性对应的用户排名设为 $N-1$ 。以此类推,将所有真实投诉用户的排名相加得到的值。

### 3.2 模型参数设置

DBN 是一类具有多层非线性映射的学习算法,它可以完成更为复杂函数的逼近,学习到更本质的刻画数据集的特征,但其对隐层相关的初始参数十分敏感。许多 DBN 应用<sup>[15]</sup>中为了获得高效的特征表示都需要调节三个参数:a) DBN 的层数;b)每一隐层的节点个数;c)训练 DBN 模型需要迭代的次数。为了获得高效的特征表示,本节将通过大量实验对比获得最优的三个参数值来设置整个模型,其中为了简化本文的模型,本文训练的 DBN 模型每层都采用相同的节点个数。由于每一隐层节点的个数和隐层的个数这两个参数是相互影响的,故本节实验时将这两个参数同时调整。对于隐层的层数本文选用了六个离散数值来进行设置,它们分别是 1、3、5、7、9、15。对于每一层隐含节点的个数本文同样选择了六个离散数值来设置,分别是 20、50、100、150、250、350、500。在评估选择这两参数时,batch\_size 的大小设置为 1 000,学习速率为 0.005,训练迭代的次数设置为 20,在微调阶段,本文使用梯度下降法微调整个网络,学习速率改为 0.1。如图 4 所示,通过同时调节不同的隐层层数和每层隐含节点的个数来训练 DBN 模型,将获得的特征向量输入随机森林(RF)后获得 AUC 评价指标。可以发现当固定每一层节点个数时,随着隐层层数不断增加,所有通过 RF 获得的 AUC 指标的曲线图几乎都是一个凸线图,而且几乎所有凸线图的 AUC 峰值都集中在隐层层数为 5 的时候。而当隐层层数固定不变来调节隐层节点数时,当隐层节点数调节至 150 时,几乎所有的 AUC 值都是最好的。综上所述,通过大量实验对比在训练 DBN 时,隐层的层数设置为 5,每层节点数设置为 150 时获得的高层特征效果最佳。

DBN 模型训练的迭代次数也是构建高效模型的有效参数。在特征构建过程中,DBN 通过调节权重  $w$  来使得原始输入特征和重建的输入特征之间的均方根误差(root mean square error, RMSE)最小。通常情况下,DBN 模型迭代的次数越多, RMSE 就越小。然而随着迭代次数的增加,模型消耗的时间就越来越长。为了平衡 RMSE 与消耗时间,本文同样选择了六个离散的迭代次数,分别是 1、5、10、20、30、50。此时本文设置层数为 5 层,每层 150 个节点,batch\_size 的大小设置为 1 000,学习速率为 0.005,在微调阶段使用梯度下降法微调整个网络,学习速率改为 0.1。利用 RMSE 和消耗时间来评价迭代次数这个参数,如图 5 结果所示,随着迭代次数的增加, RMSE 会急剧的下降,之后趋于平滑的状态,而消耗时间会随着迭代次数的增加呈倍数级的增长。根据上述的实验结果,本文最终将迭代次数设置为 15,此时的 RMSE 为 0.163 8,模型训练的时间为 27 min。

综上所述,最终 DBN 模型的层数设置为 5 层,每层节点数为 150,迭代次数为 15,batch\_size 设置为 1 000,学习速率为 0.005,在微调阶段的学习速率改为 0.1。通过训练设置最优参数的 DBN 模型来获得高阶有效的非线性组合特征表示,并将

其输入到随机森林(RF)和逻辑回归分类器(LR)中。

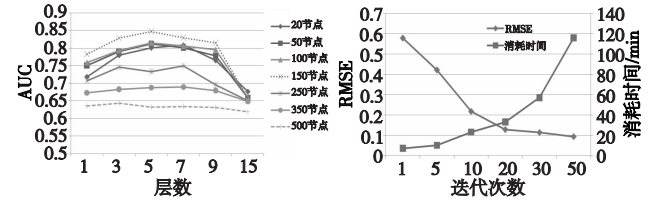


图4 不同层和节点的效果图

图5 不同迭代次数下的模型效果图

### 3.3 实验结果与分析

利用 3.2 节设置优化好的 DBN 模型参数,将原始特征输入到 DBN 模型中去构造高阶非线性组合特征,并将它们输入到随机森林(RF)和逻辑回归(LR)分类器中。如图 6 所示,将基于 DBN 特征学习的两种模型预测(DBN + RF、DBN + LR)、加入文献[4]中的学习特征(图特征和二阶组合特征)的模型预测(Learning + RF)、原始未加入学习特征的模型预测(Unlearning + RF)进行了五折交叉验证对比。首先可以发现基于 DBN 特征学习的模型和加入学习特征的模型明显好于未加入学习特征的模型预测,分别在 AUC 指标上比未加入学习特征的模型提升了 5.2%、7.1%;基于 DBN 的两种分类器预测结果都大致相同,没有过多明显的优势;最后可以发现,基于 DBN 的预测模型在 AUC 指标上比加入学习特征的模型也提升了 1.7%;总体上基于 DBN 的预测结果明显好于文献[1-3]中传统人工特征工程的结果。经过分析得到基于 DBN 预测模型能有如此良好的性能主要有以下几个原因:a)从数据输入特征的性质来看,本文 BSS 特征主要包括客户关系数据、数据采集数据、计费帐务和营销支撑数据,而 OSS 特征主要包含网络管理、网络优化方面的数据。总体来说 BSS 特征是面向运营商和服务,而 OSS 特征是面向技术支持管理。OSS 特征表现得更为底层,BSS 特征是 OSS 特征的一部分数据体现。通过 DBN 深层的网络结构使得 OSS 特征学习出了一些特征间非线性的组合关系,使获得特征更加符合投诉预测模型,让分类器更容易作出抉择。b)通过 DBN 深层的网络结构使得 BSS 特征和 OSS 特征结合得更加容易,两类特征组成了更为高阶的组合特征。

基于 DBN 模型自适应学习得到的高阶非线性组合特征和文献[4]中的学习特征都证明了其提取特征的有效。因此本文提出了将学习特征和高层非线性组合特征混合起来进行用户投诉预测,并将混合特征输入随机森林进行分类。具体的实验结果如图 7 所示。

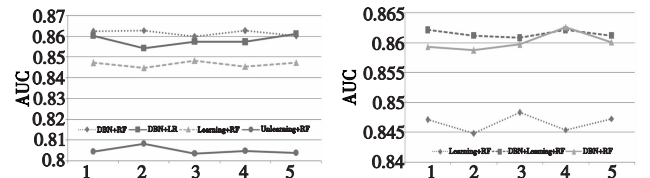


图6 四种方法的五折交叉验证图

图7 混合特征的五折交叉验证图

其中, Learning + RF 为加入学习特征的模型, DBN + RF 为利用原始特征输入 DBN 模型后将输出的高层特征输入随机森林的模型, DBN + Learning + RF 为将 34 维学习特征与通过 DBN 获得的 150 维高层特征混合输入随机森林的模型。可以看出 DBN + Learning + RF 模型比 DBN + RF 模型在 AUC 指标提升的程度,相比 DBN + RF 模型比 Learning + RF 模型在 AUC 指标提升的程度弱了很多。

通过上述理论和实验可以得到以下结论:a)本文基于深度学习的投诉预测模型通过多层神经网络获得了适合分类器分类的高层非线性组合特征,使得本文的模型精度比以往的预测模型有了很大的提升。b)本文提取有效特征的方法相比传统人工特征方法节省了大量的人力物力。c)根据 DBN + Learning + RF 模型相比 DBN + RF 模型在 AUC 性能相差无几,可以得出由 DBN 自动学习的特征也能很好地表达人工设计的学习特征。这充分表明了深度学习深层网络结构在电信大数据也能表现良好的特征学习性能。

#### 4 结束语

随着移动互联网技术的快速发展,用户对运营商产品服务的要求和对维权意识的不断提高,因此电信运营商的投诉量呈猛增趋势。为了降低用户投诉率,本文提出了基于深度学习的用户投诉预测模型,利用深度学习中 DBN 深层网络模型自动学习构造有效的高层非线性组合特征,使得预测模型在以往模型的基础上提升了 7.1%,相比传统人工特征工程方法提取特征节省了大量的人力物力。该项研究还从侧面证实了深度学习在电信大数据领域的可用性。但本文研究的问题还有许多地方需要改进,如 DBN 模型每层隐层节点数可以设置不同的值,如预测模型可以进行投诉原因多分类预测,这些问题都有待继续研究。

#### 参考文献:

- [1] 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.
- [2] 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, 5(8): 29-32.
- [3] 梁媛媛, 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.
- [4] 周文杰, 杨璐, 严建峰. 大数据驱动的投诉预测模型[J]. 计算机科学, 2016, 6(7): 53-59.
- [5] 许继元. 电信运营支撑系统[D]. 哈尔滨: 哈尔滨工程大学, 2004.
- [6] 刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. 计算机应用研究, 2014, 31(7): 1921-1930.
- [7] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. *Neural Computation*, 2006, 18(7): 1527-1554.
- [8] Shimada T, Akita K. Business support system: US, US6868390 B1[P]. 2000.
- [9] Chambers D W. Key performance indicators[J]. *Journal of the American Dental Association*, 2013, 144(3): 242-244.
- [10] Simundic A. Quality indicators[J]. *Biochemia Medica*, 2008, 18(3): 311-319.
- [11] Breiman L. Random forest[J]. *Machine Learning*, 2001, 45(1): 5-32.
- [12] 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.
- [13] Bergstra J, Breuleux O, Bastien F, et al. Theano: a CPU and GPU math compiler in python[C]//Proc. 9th Python in Science Conference. 2010: 1-7.
- [14] Guyon I, Lemaire V, Boule M, et al. Analysis of the KDD cup 2009: fast scoring on a large orange customer database[C]//Proc of KDD. 2009: 8.
- [15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. *Advances in Neural Information Processing Systems*, 2012, 25(2): 2012.
- [16] HBase 查询技术与系统[J]. 计算机学报, 2016, 39(1): 140-153.
- [17] Christian S J, Dan L, Beng C O. Query and update efficient B + -tree based indexing of moving objects[C]//Proc of the 30th International Conference on Very Large Data Bases. Toronto, Canada: VLDB Endowment, 2004.
- [18] Chen N, Shou L D, Chen G, et al. B(S)-tree: a self-tuning index of moving objects[C]//Proc of Database System for Advanced Applications. 2010: 1-16.
- [19] Norbert B, Hans-Peter K, Ralf S, et al. The R\*-tree: an efficient and robust access method for points and rectangles[J]. *SIGMOD Rec*, 1990, 19(2): 322-331.
- [20] Choi Y J, Min J K, Chung C W. A cost model for spatio-temporal queries using the TPR-TREE[J]. *J SystSoftw*, 2004, 73(1): 101-112.
- [21] Kim S W, Jang M H, Lim S. Active adjustment: an effective method for keeping the TPR\*-tree compact[J]. *Journal of Information Science and Engineering*, 2010, 26(5): 1583-1600.
- [22] 郑祖芳. 分布式并行时空索引技术研究[D]. 武汉: 中国地质大学, 2014.
- [23] 张榆, 马友忠, 孟小峰. 一种基于 HBase 的高效空间关键字查询策略[J]. 小型微型计算机系统, 2012, 33(10): 2141-2146.
- [24] 万静, 张义, 何云斌, 等. 基于 KD-树和 K-means 动态聚类方法研究[J]. 计算机应用研究, 2015, 32(12): 3590-3595.
- [25] 迪米达克. HBase 实战[M]. 北京: 人民邮电出版社, 2013.
- [26] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.
- [27] 李德仁, 马军, 邵振峰. 论时空大数据及其应用[J]. 卫星应用, 2015, 11(9): 7-11.
- [28] Li Zhisheng, Lee K C K, Zheng Baihua, et al. IR-tree: an efficient index for geographic document search[J]. *IEEE Trans on Knowledge and Data Engineering*, 2011, 23(4): 585-599.
- [29] 吴钦阳. R\*-树空间索引的改进[J]. 计算机应用, 2010, 30(2): 419-422.
- [30] 孙冬璞, 郝晓红, 郝忠孝. 频繁更新移动对象的索引方法[J]. 计算机工程, 2013, 39(11): 52-56.
- [31] 汪璟玢. 一种结合空间聚类算法的 R 树优化算法[J]. 计算机工程与应用, 2014, 50(5): 112-115.
- [32] 王龙. 基于 Hilbert 曲线的查询方法研究[D]. 哈尔滨: 哈尔滨理工大学, 2013.
- [33] 陈新鹏. 基于 HBase 的数据生成与索引方法的研究[D]. 北京: 北京邮电大学, 2014.
- [34] Vashishtha H, Stroulia E. Enhancing query support in HBase via an extended coprocessors framework[C]//Proc of the 4th European Conference on Towards a Service-Based Internet. 2011: 75-87.
- [35] 葛微, 罗圣美, 周文辉, 等. HiBase: 一种基于分层式索引的高效

(上接第 页)用索引表进行辅助查询,用以提高时空关联查询的效率。

#### 参考文献:

- [1] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9): 1889-1908.
- [2] 李德仁, 马军, 邵振峰. 论时空大数据及其应用[J]. 卫星应用, 2015, 11(9): 7-11.
- [3] Li Zhisheng, Lee K C K, Zheng Baihua, et al. IR-tree: an efficient index for geographic document search[J]. *IEEE Trans on Knowledge and Data Engineering*, 2011, 23(4): 585-599.
- [4] 吴钦阳. R\*-树空间索引的改进[J]. 计算机应用, 2010, 30(2): 419-422.
- [5] 孙冬璞, 郝晓红, 郝忠孝. 频繁更新移动对象的索引方法[J]. 计算机工程, 2013, 39(11): 52-56.
- [6] 汪璟玢. 一种结合空间聚类算法的 R 树优化算法[J]. 计算机工程与应用, 2014, 50(5): 112-115.
- [7] 王龙. 基于 Hilbert 曲线的查询方法研究[D]. 哈尔滨: 哈尔滨理工大学, 2013.
- [8] 陈新鹏. 基于 HBase 的数据生成与索引方法的研究[D]. 北京: 北京邮电大学, 2014.
- [9] Vashishtha H, Stroulia E. Enhancing query support in HBase via an extended coprocessors framework[C]//Proc of the 4th European Conference on Towards a Service-Based Internet. 2011: 75-87.
- [10] 葛微, 罗圣美, 周文辉, 等. HiBase: 一种基于分层式索引的高效