深度随机森林在离网预测中的应用

杨晓峰[1] 严建峰[1,2] 刘晓升[1] 杨 璐[1]

(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 在电信运营商领域,离网预测模型是企业决策者用来发现潜在离网用户(即停用运营商服务)的主要手段。目前离网预测模型都是基于逻辑回归、决策树、神经网络及随机森林等浅层机器学习算法,但是在大数据的背景下,这些浅层算法在预测问题上很难取得更高的精度。因此,提出了一种新型的深层结构模型——深度随机森林,通过将传统浅层随机森林堆积成深层结构模型,获得更高的预测精度。在运营商真实数据上进行了大量实验,结果证明深层随机森林模型比传统浅层机器学习算法在离网预测问题上可以得到更好的效果。同时,增大训练数据量可以进一步提升深层随机森林的预测能力,从而证明了在大数据环境下深层模型的潜力。

关键词 离网预测,深层随机森林

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.042

Deep Random Forest for Churn Prediction

YANG Xiao-feng[1] YAN Jian-feng[1,2] LIU Xiao-sheng[1] YANG Lu[1]

(College 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 Churn prediction models help telecom operators identify potential off-network user. Most previous models adopt shallow machine learning algorithms such as logistic regression, decision tree, random forest and neural networks. This paper proposed a novel deep random forest algorithm, which is a multi-layer random forest with layer-wise training. In terms of telecom operators' real data, we confirmed that the proposed deep random forest performs better than previous shallow learning algorithms in churn prediction. Moreover, increasing the volume of training data can further improve the performance of deep random forest, which implies that big data make deep models advantageous over shallow models.

Keywords Churn prediction, Deep random forest

1 引言

用户流失是目前很多运营商都面临的严重问题。随着电信运营行业的蓬勃发展,各企业之间的竞争也越来越激烈。很多其它运营商通常会提供一些优惠政策来吸引用户,这是许多用户离网的最主要原因。在电信运营行业中,基于付费方式的不同通常将用户分为预付费和后付费两类,所谓后付费用户是指与运营商签订协议的用户,而其他用户则统称为预付费用户。相比而言,预付费用户不够稳定,他们随时都有离网的可能。图1给出了上海某运营商2014年连续10个月的离网率分布情况,从图1中可以了解到预付费用户每个月的离网率达到10%左右,而后付费用户只有5%左右。因此,通过离网预测的方式来挽留预付费用户是目前运营商决策者解决用户流失问题的主要前提手段。

离网预测[9,10]的目的是让决策者提前预判潜在的离网用户。现在,运营商决策者通常会运用数据挖掘和统计工具来判别未来可能离网的用户。运营商往往会拥有大量用户数据,包括账单信息、通话详单和网络数据等。通过系统地分析这些历史数据,可以挖掘出很多隐藏在这些数据之后的信息,进而可以更好地预判用户下一步的行为。

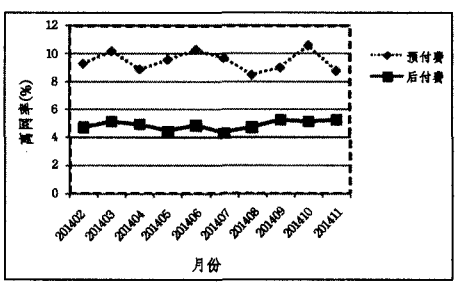


图1 2014年连续10个月的离网率

离网预测包含两个步骤,即构造有用的特征和建立一个

到稿日期:2015-03-27 返修日期:2015-07-10 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),广东省重点实验室开放课题(SZU-GDPHPCL-2012-09)资助。

杨晓峰(1990—),男,硕士生,主要研究方向为机器学习;严建峰(1978—),男,副教授,硕士生导师,主要研究方向为机器学习,E-mail:yanjf@suda.edu.cn(通信作者);刘晓升(1976—),男,博士生,主要研究方向为机器学习;杨 璐(1982—),女,副教授,硕士生导师,主要研究方向为机器学习与软件工程。

良好的分类器。首先从若干子表里面抽取有用特征组成大宽表作为训练集及测试集，然后选择分类算法训练带标签的训练数据集得到分类器。学者在之前使用过不同的分类算法来处理离网预测问题，例如逻辑回归[6-8]、决策树[8,11]、boosting算法[13]、随机森林[9,14]、神经网络[15,16]及支持向量机[17]等。但是这些模型在处理电信大数据时都存有弱点，因为它们都是属于浅层机器学习算法，而浅层机器学习方法对数据量有一定的局限性[12]。当数据量大到一定程度时，浅层算法学习能力的就不如深层结构算法的强。这是其模型自身特点及大数据背景所决定的。

深度学习是目前在工业界和学术界都非常火热的深度结构机器学习算法。深度学习成功的两个基本条件分别是大数据的时代背景和高性能的硬件设备。但由于实验设备的限制，深度学习方法在实际中训练时间太长，而且参数调制复杂，因此并没有在离网预测模型中直接应用。

本文提出了一种新型的深度结构算法——深度随机森林，它是基于深度学习结构，将随机森林堆积成多层结构的深度模型。选择随机森林作为基础算法最主要的原因是其训练速度快，训练多层随机森林也不会非常耗时。通过实验发现，本算法在离网预测模型中的预测精度比浅层学习算法高。

2 深度学习和随机森林

2.1 深度学习的概念

深度学习的概念起源于对人工神经网络的研究，于2006年由Hinton等人正式提出[1]。基于深信度网(DBN)提出非监督贪心逐层训练算法，为解决深层结构相关的优化难题带来希望，随后多层自动编码器深层结构被提出。此外Lecun等人提出的卷积神经网络[2,18]是第一个真正的多层结构学习算法，它利用空间相对关系减少参数数目以提高训练性能。深度学习结构通常包含一个输入层、多个隐层以及一个输出层。具有多个隐含层的多层感知器是深度学习模型的一个典型例子。

2.2 深度学习的原理

深度学习的原理可作如下描述：假设有一个系统S，它是n层结构模型，记为$(s_1,\cdots,s_i,\cdots,s_n)$，其中$s_i$表示第$i$层，且输入记为$I$，输出记为$O$，该模型也可以表示为：$I\Rightarrow s_1\Rightarrow s_2\Rightarrow\cdots\Rightarrow s_n\Rightarrow O$，如果输出$O$等于输入$I$，即输入$I$经过这个系统变化之后没有任何的信息损失，保持不变，即输入I经过每一层s_i都没有任何的信息损失，也可以认为任何一层s_i都是原始输入I的另外一种表示。所以深度学习的最大特点就是自动地学习特征。假设有一堆输入I(如一堆图像或者文本)，且有一个n层的网络系统S，可以通过调整这个网络结构系统的参数，使得其输出仍然是输入I，那么就可以自动地获取得到输入I的一系列层次特征。对于深度学习来说，其结构就是堆叠多个层，将前一层的输出作为后一层的输入。深度学习的实质是通过构造多层的复杂学习模型和海量的数据样本来学习更有用的特征，进而提升分类和预测的准确性。深度学习通过复杂的高容量模型可以有效地学习大数据内部隐藏的复杂多变的高阶统计特性。

2.3 随机森林的概念

决策树是一种普遍使用的具有很多良好特性的分类算法，它的主要特点有训练时间复杂度低、预测的过程比较快、模型容易展示等。但是单决策树又有一些不足的地方，比如容易过拟合，虽然可以通过剪枝等方法减少这种情况的发生，但仍有不足。2001年Leo Breiman[3]在决策树的基础上提出了随机森林算法。随机森林是由多棵依赖相同分布的决策树组成的一个分类器，并且其输出的类别是由个别树输出的类别的众数而定。

2.4 随机森林的原理

随机森林[3,4]是一种统计学习理论，它利用bootstrap重采样方法从原始样本中抽取多个样本，并对每个bootstrap样本进行决策树建模，然后综合多棵决策树的预测结果，通过投票或者取平均等方式得到最终预测结果。其也可以看作是将若干个弱分类器组合成一个强分类器。

随机森林通过选取不同的训练集及特征集来增加分类模型间的差异，从而提升组合分类模型的外推预测能力。通过K次训练，得到K棵不同的决策树$\{T_1,T_2,\cdots,T_K\}$，再将这些树组合成一个分类模型系统，该系统的最终分类结果大多采用简单的多数投票法，可以表示成：

$$H(x)=\arg\max_{Y}\sum_{i=1}^{K}I(h_i(x)=Y) \tag{1}$$

其中，$H(x)$表示最后分类结果，$h_i(x)$表示单棵决策树的分类结果，Y表示类别标签，$I(\cdot)$则表示示性函数。

3 模型框架

整个离网预测模型一共包含5个组成部分，分别是源数据处理、特征提取、制定标签、训练模型分类器以及评价分类性能。图2展示了整个模型的结构框架。

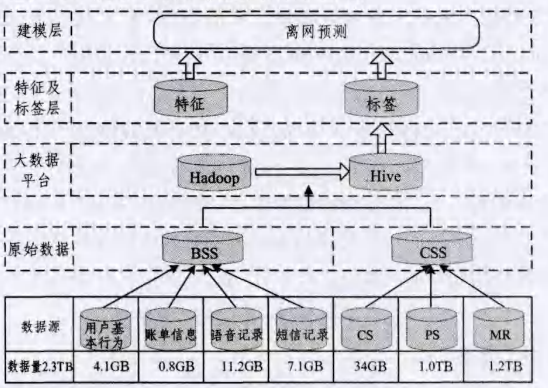


图2 离网预测模型框架

3.1 大数据平台

本文使用的数据来源于基于上海某运营商的大数据平台。该数据平台每天都会产生将近2.3TB大小的数据，包括BSS(Business Support System)数据及OSS(Operation Support System)数据，其中OSS数据总量占比约97%左右。BSS又称为业务支持系统，通常BSS数据包含有用户基本信息、用户行为、账单信息、语音数据、短信数据及通话详单等，大概每天产生的数据量在24GB左右。而OSS全称是运营支持系统，它的数据主要包含3块，分别是CS(Circuit Switch)数据、PS(Packet Switch)数据及MR(Measurement Report)数据，其中CS数据是指通话连接质量数据，例如通话掉话率和通话连接成功率等；PS数据指用户手机上网行为数据，包括手机上网速度、连接网络成功率和移动搜索等；而MR数据采自无线网络控制器，可以描述用户的大致轨迹信息。整个

OSS 每天产生的数据量约为 2.2TB。

BSS 数据可以理解为与用户密切关联的数据，它内部有 140 张表；而 OSS 数据则是网络服务层的数据，包括用户流量详细记录表、交易详细记录表、统计详细记录表等，这些表可以通过国际移动用户识别码这一共同的主键进行关联。在数据管理这一层面，选择使用 Apache Hadoop 分布式结构作为管理和使用数据的技术支撑。Hadoop 的分布式文件系统 HDFS 可以处理 PB 级别的超大文件，而且可以流式地访问数据，更重要的是它可以部署在普通的商用机器集群上；同时 Hadoop 的数据仓库工具 Hive 可以提供简单的 SQL 查询功能，并且可以将 SQL 语句转化为 MapReduce 任务分布式运行，容易实现并行化。

3.2 特征工程

离网预测的特征工程是基于 Hive/Spark SQL 及一些广泛使用的非监督/监督学习算法来完成的，包括 PageRank[20]、Label Propagation[21] 和主题模型[23] 等。项目涉及的所有原始数据都以 Hive 表的形式存储到 HDFS 里面，然后用 Hive SQL 或 Spark SQL 的关联和聚合操作将部分有用特征字段做成临时表，这些临时表可以重复使用，最后将所有的有用特征字段汇成一张大宽表，表内每一行向量就表示一条用户数据。

离网预测模型涉及到的特征主要分为 5 块：基本特征、CS 特征、PS 特征、基于图的特征及二次特征。基本特征主要是从 BSS 数据中抽取，包括账单、余额、通话频率、通话时长、投诉频率、充值金额等。而 CS 及 PS 特征则是从 OSS 得到。CS 特征指的是用户通话质量方面的特征，主要包括 Uplink MOS、Downlink MOS、IP MOS 等，它们可以评估用户语音服务的质量；而 PS 特征是指网络服务质量，主要涉及上网、收发邮件及流媒体等服务情况数据。基于图的特征则是通过 PageRank 和 Label Propagation 算法从通话图、短信图中抽取出来的特征。而二次特征则是通过 LIBFM[22] 将用户影响力特征排名前 5 位的特征两两相乘而得。最后将所有的特征拼接成特征向量，每个用户就可以用 $X_m=[x_1,\cdots,x_i,\cdots,x_j,\cdots,x_N]$ 来表示，其中 x_i 表示用户 X_m 的第 i 个特征。

3.3 标签

与特征一样，标签也是数据挖掘过程中不可或缺的组成部分，标签设定的好坏可直接影响整个模型的优劣。在离网预测模型中，需要标注出离网用户和在网用户，由于没有一个明确的字段表明用户是否离网，因此需要通过行业经验和常识去判断。最后打标签的规则设为将进入充值期后且 15 天不充值的用户标注为离网用户，其他则标注为在网用户，其中充值期是指余额小于 0 的这段时间。因此预测用户下个月是否离网就可以转化成预测用户是否下个月进入充值期后且在 15 天内不充值的问题。表 1 列出了过去 9 个月用户进入充值期后进行充值的比例分布。从表 1 中可以看出，只有 7% 的用户会在进入充值期后 15 天内不充值，说明这样的情况并不占多数。

表 1 用户进入充值期充值时间分布

天数	0～5	6～10	11～15	15 以上
充值人数比例(%)	65	19	9	7

3.4 分类器

随机森林是目前工业界相当火热的分类算法，训练速度快及建模精度高等特点是其目前被广泛应用的主要原因。但是我们之前也讨论过，在处理大数据时随机森林等浅层机器学习方法的分类性能不如深层结构算法。深层结构机器学习方法(比如深度学习)可以通过构造多层的复杂结构模型来学习大数据内部隐藏的复杂多变的高阶统计特性。大数据量和高性能的硬件设备是深度学习成功的两个必备条件。但由于实验环境限制，深度学习在我们的模型中并不适用，其突出问题是训练时间长、参数调整复杂。

针对以上所述，本文提出一种新型的深度结构模型，称作深度随机森林，它是基于深度学习结构将随机森林堆积成多层结构的深度模型。选择随机森林作为我们的基础算法的最主要原因是其具有训练速度快的特点，在训练多层随机森林时不会非常耗时。图 3 给出了深度随机森林的结构图。

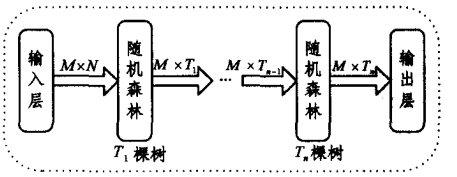


图 3 深度随机森林的结构图

在机器学习中，标签是最好的特征。如果将标签作为特征去训练分类器，学习出来的分类模型的预测精度一定非常高。因此一方面为了提升模型预测精度，必须找到与标签密切相关的特征；另一方面，分类器每次的输出结果就是预测标签。鉴于这两点，可以将分类器的输出结果作为训练集输入到另一个分类器中进行训练，相当于将学习到的预测标签作为特征。这个思想借鉴了深度学习的思想。深度学习的实质是从复杂数据结构里面自动学习与标签密切关联的特征，再用学习到的特征去预测标签。而本文提出的深度随机森林，是将前一层学习到的预测标签作为下一层的特征输入。所以两者都是通过学习到的特征去预测标签。

整个模型的设计流程可以描述如下：假设深度随机森林是 $n(n\geqslant 4)$ 层结构模型，分别是输入层、随机森林层及输出层，其中随机森林层一共有 $n-2$ 层。给定一组训练样例 $\{x_m\}$，样本总数为 M，而 x_m 表示第 m 个用户，类标签记为 y_m，其中离网用户用数值 1 表示，在网用户则用 0 表示。首先训练第一层随机森林，输入是 $M\times N$ 的矩阵，训练结束后再将第一层的输出作为第二层的输入来训练第二层随机森林，比如第一层有 T_1 棵树，那么每个实例 x 在通过第一层后每棵树都会给出其属于正例的概率值，这样每个实例都会有 T_1 个输出值，然后将这 T_1 个值处理成向量作为第二层随机森林的输入，即输入是 $M\times T_1$ 的矩阵，而类别标签仍然不变，继续训练第二层，训练结束后将第二层的输出作为第三层的输入再进行训练，依此类推，直到全部训练结束。而在测试阶段，同样先将测试样例输入到第一层，再将输出结果输入到下一层，再依次往下，到最后输出层时，可以通过式(2)得到测试样本为正例的概率值。

$$y=\frac{1}{T_n}\sum_{i=1}^{T} f_i(x) \tag{2}$$

其中，y 是样本 x 属于离网的概率值，$f_t(\cdot)$ 是每棵树给出的分类结果，T_n 则表示树的总棵数。

提出的深度随机森林算法与 stacking 算法有一定联系。这两个算法都是多层学习模型，而且都是将第一层若干个弱分类器的输出结果作为第二层的输入。但这两个算法也存在诸多不同的地方。首先，stacking 算法是两层学习模型，而提出的深度随机森林算法是多层学习模型，它的层数理论上并不受限制。其次，stacking 算法在训练第一层的每一个弱分类器时都用的是全量训练数据，而深度随机森林在训练每一层时采用 bootstrap 重采用方法。

4 实验

在本文实验中，每次实验需要一组连续 4 个月的数据，具体参照图 4。首先用第 N 个月的标签匹上第 $N-1$ 个月的特征作为训练数据放进模型里进行训练。然后再将无标签的第 N 个月的特征作为测试数据输入到已经学习好的分类器里面，预测第 $N+1$ 个月的离网用户。最后将第 $N+2$ 个月给第 $N+1$ 个月所打的标签与我们的预测名单进行对比来评估模型的性能指标。

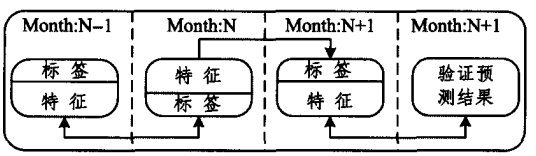


图 4 数据处理流程

在实验中，为了确定每层森林包含树的棵数，用单层随机森林进行相关实验，具体如图 5 所示。

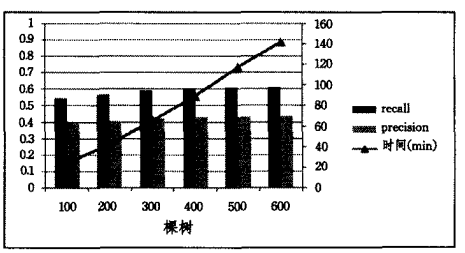


图 5 树的棵树对训练时间及模型预测结果的影响

从图 5 中可以看到，训练时间随着树的棵数变化基本呈线性增长趋势，但模型的预测结果在 400 棵树之后基本保持平稳，故选择 400 棵树作为每一层的树的棵数。

4.1 评价指标

使用查全率(Recall)和查准率(Precision)作为模型的评价指标。离网预测模型每次都会根据输出概率降序选择前 U 位用户作为预测名单，通常若 U 越大，则查全率越高，查准率越低。因此，选择合适大小的 U 对得到满意的查全率和查准率至关重要。式(3)和式(4)分别对应离网预测模型中查全率及查准率的定义：

$$Recall=\frac{U_{True}}{Total_{True}} \tag{3}$$

$$Precision=\frac{U_{True}}{U} \tag{4}$$

其中，U_{True} 表示在选择的 U 位用户里面真正离网的用户数，$Total_{True}$ 表示在预测的当月真正离网的用户数。除此之外，也选用了 AUC 和 PR-AUC[5] 作为评价指标，其中 AUC 可以定义成：

$$AUC=\frac{\sum_{c\text{为真正离网用户}} rank_c-\frac{P\times(P+1)}{2}}{P\times N} \tag{5}$$

其中，P 表示真正的离网用户数，N 表示真正的在网用户数，$rank_c$ 表示用户 c 在给出的离网概率中的排名，其中概率值最高的赋值为 n，次高的则为 $n-1$，依此类推。AUC 值越高则说明模型越好，但是考虑到在我们的样本中正负样例的比例严重不平衡，负样例数只占总数的 10%左右，因此又选用 PR-AUC 作为衡量标准，它在不平衡数据上的表现比 AUC 更出色。在此基础上，还选择用卡方检验方法[24]进一步验证提出的方法的有效性。

4.2 深度随机森林与随机森林对比

在本节实验中，主要是对深度随机森林模型与单层随机森林进行比较，并通过改变深度随机森林的层数来验证模型结构的深度对分类效果的影响。

具体实验内容是用 6 月份的特征关联 7 月份的标签作为训练数据，然后用 7 月份的特征作为测试数据预测 8 月份的离网用户，并且根据输出的离网概率降序取前 30 位用户作为预测名单。图 6 给出深度随机森林与单层随机森林的实验结果对比。

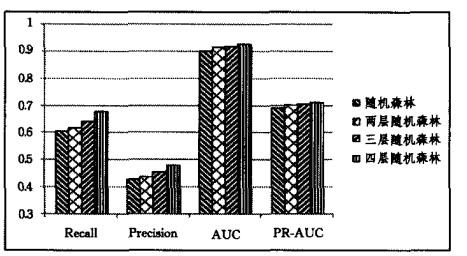


图 6 深度随机森林与单层随机森林实验对比

从图 6 中可以看出，深度随机森林在各项评价指标上相对于单层随机森林都占有明显优势。而且可以发现，深度随机森林的层次越深，模型表现出来的分类性能越好。

在此基础上，还选择用卡方检验方法验证了深层随机森林相对于单层随机森林算法更有效，具体如表 2 所列。

表 2 深层随机森林与单层随机森林对比的卡方检验结果

对比的模型	两层随机森林	三层随机森林	四层随机森林
χ^2	61.0218	472.926	1672.66
P	5.6E−15	7E−105	0
是否存在显著性差异(99%的置信水平)	存在	存在	存在

从表 2 中可以看出，在 99%的置信水平下，深层随机森林相对于单层随机森林存在显著性差异，而且随着层数增多，差异也越大。故进一步验证了本文提出的深层随机深林更加有效。

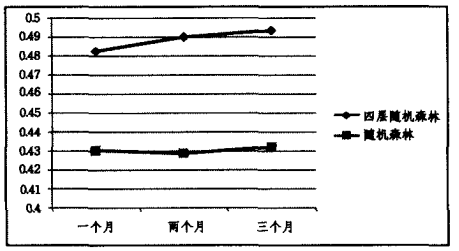
4.3 深度随机森林与数据量的关系

本节实验主要验证数据量对深度随机森林的影响。首先选用 5 月、6 月的特征及其对应标签作为训练数据放进四层的深度随机森林中进行训练，之后预测 8 月份的离网用户；然后再扩大数据量，用 4−6 月这 3 个月的数据作为训练数据，同样预测 8 月份的离网用户。实验结果如表 3 所列。

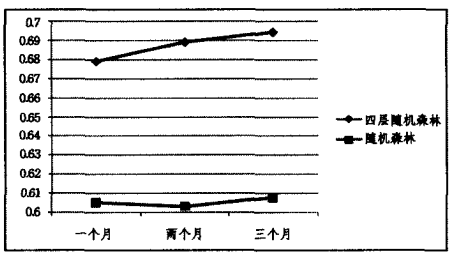
表 3　数据量与深度随机森林关系的实验对比

Evaluation Index	Recall	Precision	AUC	PR-AUC
One Month Data	0.6791	0.4829	0.9308	0.7132
Two Months Data	0.6890	0.4900	0.9352	0.7169
Three Months Data	**0.6943**	**0.4937**	**0.9360**	**0.7177**

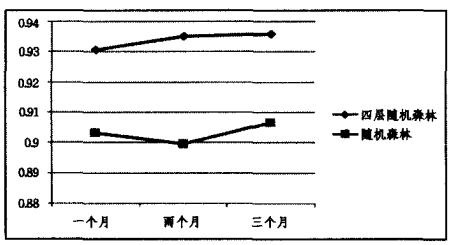
表 3 给出了 3 个不同数据量(分别对应 1 个月、2 个月及 3 个月的数据)参与训练产生的实验结果对比。可以看出,数据量越大,模型分类能力越强。但表 3 还不能充分地说明深度结构模型是使得数据量与分类效果密切关联的主要原因。图 7 给出另外一张实验对比图,它将 4 层的深度随机森林与单层随机森林在不同数据量下的实验结果进行了对比。



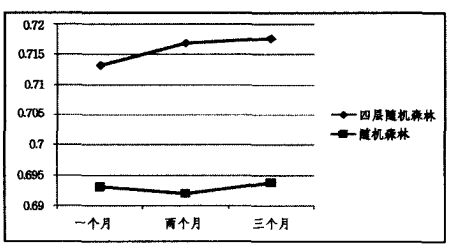
(a)Recall



(b)Precision



(c)AUC



(d)PR-AUC

图 7　深度随机森林与随机森林在不同数据量下的指标对比

图 7 表明,随着数据量的增大,四层随机森林的各项评价指标明显上升,而随机森林的评价指标保持平稳波动。因此,我们可以认为深层结构模型处理大数据分类问题的能力明显比浅层机器学习强大。

4.4　深度随机森林与其他机器学习算法对比

本节实验主要是将深度随机森林与其它浅层机器学习算法进行比对。我们选取了支持向量机、LIBLINEAR 及 GBDT 这些主流的机器学习算法作为对比对象。实现结果如图 8 所示。

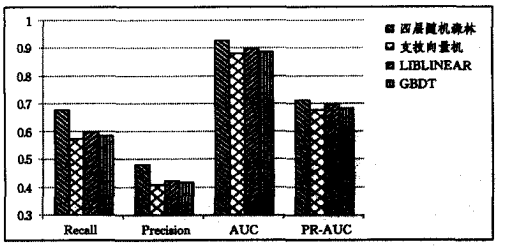


图 8　深度随机森林与主流机器学习算法对比

图 8 表明,深度随机森林的分类结果指标明显优于其它 3 个机器学习算法。由此可以得出结论,深度结构模型处理大数据分类问题比浅层结构模型更具优势。

结束语　本文将深度学习与随机森林相结合,提出一种新型的深度结构模型——深度随机森林,并成功地将其应用到离网预测模型中,得到了若干有指导意义的结论。与传统随机森林算法相比,本文提出的深度随机森林模型在处理大数据分类问题时表现得更好一些。深度结构模型是大数据背景的产物,如何构造一个通用的深层结构模型是一个值得研究的课题。本文提出的深度随机森林还不够完善,它的逐层训练模式还不够快速,以及如何调整每个森林的参数使其可以达到最好的分类效果等问题都有待进一步研究。

参 考 文 献

[1] Hinton G E, Osindero S. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554

[2] LeCun Y, Jackel L, Bottou L, et al. Comparison of Learning Algorithms for Handwritten Digit Recognition[C]// International Conference on Artifical Neural Networks. 1995: 53-60

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

[4] Fang Kuang-nan, Wu Jian-bin, Zhu Jian-ping, et al. A Review of Technologies on Random Forests[J]. Statistics and Information Forum, 2011, 26(3): 32-38(in Chinese)
方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛,2011,26(3):32-38

[5] Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves[C]// Proceedings of the 23rd International Conference on Machine Learning(ICML). 2000: 233-240

[6] Neslin S, Gupta S, Kamakura W A, et al. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models[J]. Social Science Electronic Publishing, 2006, 43(2): 204-211

[7] Hadden J, Tiwari A, Roy R, et al. Computer assisted customer churn management: State-of-the-art and future trends[J]. Computers & Operations Research, 2007, 34(10): 2902-2917

[8] Lima E. Domain knowledge integration in data mining using decision tables: case studies in churn prediction[J]. Journal of the Operational Research Society, 2009, 60(8): 1096-1106

[9] Huang Yi-qing, Zhu Fang-zhou, Yuan Ming-xuan, et al. Telco churn prediction with big data[C]// SIGMOD. 2015: 607-618

[10] Yuan Ming-xuan, Deng Ke, Zeng Jia, et al. OceanST: A distributed analytic system for large-scale spatiotemporal mobile broadband data[C]// VLDB (Demo). 2014: 1561-1564

[11] Verbeke W, Martens D, Mues C, et al. Building comprehensible

customer churn prediction models with advanced rule induction techniques[J]. Expert Systems with Applications, 2011, 38(3): 2354-2364

[12] Sun Zhi-jun, Xue Lei, Xu Yang-ming, et al. Overview of deep learning[J]. Application Research of Computers, 2012, 29(8): 2806-2810(in Chinese)
孙志军,薛磊,许阳明,等. 深度学习研究综述[J]. 计算机应用研究,2012,29(8):2806-2810

[13] S Jin-bo, L Xiu, L Wen-huang. The Application ofAdaBoost in Customer Churn Prediction[C]// 2007 International Conference on Service Systems and Service Management. IEEE, 2007: 1-6

[14] Lemmens A, Croux C. Bagging and boosting classification trees to predict churn[J]. Journal of Marketing Research, 2006, 43 (2): 276-286

[15] Datta P, Masand B R. Mani D, et al. Automated Cellular Modeling and Prediction on a Large Scale[J]. Artificial Intelligence Review, 2000, 14(6): 485-502

[16] Hung S, Yen D C, Wang H. Applying data mining to telecom churn management [J]. Expert Systems with Applications, 2006, 31: 515-524

[17] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction[J]. Dirk Van den Poel, 2008, 36(3): 4626-4636

[18] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. 1998

[19] Liu Jian-wei, Liu Yuan, Luo Xiong-lin. Research and development on deep learning[J]. Application Research of Computers, 2014, 31(7): 1921-1930(in Chinese)
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

[20] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[C]// Stanford InfoLab. 1998: 1-14

[21] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[R]. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002

[22] Rendle S. Scaling factorization machines to relational data[J]. PVLDB, 2013, 6(5): 337-348

[23] Zeng J, Cheung W K, Liu J. Learning topic models by belief propagation[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2013, 35(5): 1121-1134

[24] Xu Xiang-yang. Application of x2 Test in Analysing Students' Score Difference[J]. Journal of Changzhou Teachers College of Technology, 2001, 7(4): 13-16(in Chinese)
徐向阳. 卡方检验在学生成绩差异性分析中的应用[J]. 常州技术师范学院学报,2001,7(4):13-16

(上接第 203 页)

[36] Jia X, Xin F, Chuan W R. Adaptive spray routing for opportunistic networks[J]. International Journal on Smart Sensing and Intelligent Systems, 2013, 6(1): 95-119

[37] Sarkar P, Chakrabarti D, Jordan M I. Nonparametric Link Prediction in Dynamic Networks[C]// Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, UK, 2012

[38] He Yu-lin, Liu J N K, Hu Yan-xing, et al. OWA operator based link prediction ensemble for social network[J]. Expert Systems with Applications, 2015, 42(1): 21-50

[39] Barbieri N, Bonchi F, Manco G. Who to follow and why: link prediction with explanations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 1266-1275

[40] Bliss C A, Frank M R, Danforth C M. Peter Sheridan Dodds, An evolutionary algorithm approach to link prediction in dynamic social networks[J]. Journal of Computational Science, 2014, 5 (5): 750-764

[41] Cao Bin, Liu Nan, Yang Qiang. Transfer Learning for Collective Link Prediction in Multiple, Heterogenous Domains[C]// Proceedings of the International Conference on Machine Learning. 2010: 159-166

[42] Melville P, Sindhwani V. Recommender systems[C]// Encyclopedia of Machine Learning. Springer, 2010

[43] Bartunov S, Korshunov A, Park S T, et al. Joint link-attribute user identity resolution in online social networks[C]// Proceedings of the Workshop on Social Network Mining and Analysis (SNA-KDD). 2012

[44] Bliss C A, Frank M R, Danforth C M. Peter Sheridan Dodds, An evolutionary algorithm approach to link prediction in dynamic social networks[J]. Journal of Computational Science, 2014, 5 (5): 750-764

[45] Li X, Chen H C. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach[J]. Decision Support Systems, 2013, 54(2): 880-890

[46] Zeng Z Z, Chen Ke-jia, Zhang Shao-bo, et al. A link prediction approach using semi-supervised learning in dynamic networks [C]// 2013 Sixth International Conference on Advanced Computational Intelligence (ICACI). 2013: 276-280

[47] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks[C]// Proceedings of the WSDM Conference. 2010: 635-644

[48] Menon A KElkan C. Link prediction via matrix factorization[C]// Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). 2011: 437-452

[49] Hu F Y, Wong H S. Labelling of Human Motion Based on CB-GA and Probabilistic Model[J]. International Journal on Smart Sensing and Intelligent Systems, 2013, 6(2): 583-609

[50] Jia X, Xin F, Chuan W R. Adaptive spray routing for opportunistic networks[J]. International Journal on Smart Sensing and Intelligent Systems, 2013, 6(1): 95-119

[51] Kang C, Pugliese A, Grant J, et al. STUN: querying spatio-temporal uncertain(social) networks[J]. Social Network Analysis & Mining, 2014, 4(1): 1-19

[52] Backstrom L, Leskovec J. Supervised random walks Predicting and recommending links in social networks[C]// Proceedings of the WSDM Conference. 2010: 635-644

[53] Luong N T, Nguyen T T, Jung J J, et al. Discovering Co-author Relationship in Bibliographic Data Using Similarity Measures and Random Walk Model[J]. Intelligent Information and Database Systems, Lecture Notes in Computer Science, 2015, 9011: 127-136