基于蜂窝网络实测数据的电信诈骗检测模型研究

郭开远，王文博
（北京邮电大学信息与通信工程学院，北京 100876）

摘要：随着无线通信技术的迅猛发展，利用手机电话等通讯手段进行电信诈骗，已经成为危害用户安全的一大问题。针对这一问题，论文通过对蜂窝网络数据进行深入分析与挖掘，构建电信诈骗用户检测模型，该模型包括数据预处理，CNNcombine 算法、模型评估三部分。首先在数据处理部分，对数据集进行了特征筛选、编码、抽样等工作。其次，CNNcombine 算法是本文提出的一种一维卷积神经网络与多个传统分类算法结合的算法，将卷积神经网络运用于解决文本像信号以外的分类问题。最后在模型评估部分，证明了 CNNcombine 算法比 XGBoost 等常用机器学习分类算法检测电信诈骗用户上有更高的准确率。

关键词：机器学习；蜂窝网络数据；深度学习；分类算法
中图分类号：TP181

Research on Telecom Fraud Detection Model Based on Cellular Network Data

GUO Kaiyuan, WANG Wenbo
(School of Information and Communication Engineering,Beijing University of Posts and Telecommunications,Beijing 100876)

Abstract: With the rapid development of wireless communication technology, the use of mobile phones and other means of communication for telecommunications fraud has become a major problem that endangers user security. Aiming at this problem, this paper constructs a telecom fraud user detection model by in-depth analysis and mining of cellular network data. The model includes data processing, CNNcombine algorithm and model evaluation. First, in the data processing part, the data set is subjected to feature screening, coding, sampling, and the like. Secondly, the CNNcombine algorithm is a combination of a one-dimensional convolutional neural network and multiple traditional classification algorithms. The convolutional neural network is applied to solve classification problems other than text image signals. Finally, in the model evaluation part, it is proved that the CNNcombine algorithm has higher accuracy than the common machine learning classification algorithm such as XGBoost to detect telecom fraud users.

Key words: Machine learning; cellular network data; deep learning; classification algorithm

0 引言

“电信诈骗”是指犯罪分子通过电话语音、短信等手段，发送虚假信息，设置诈骗骗局，以及对受害者实施远程非接触式诈骗，诱使受害者支付或转账给犯罪分子的犯罪行为，长期影响着社会的安全和稳定。电信运营商拥有海量的数据，数据规模远远大于其他行业的统计数据，本文的研究重点如何对蜂窝网络数据进行分析并建模，并利用这些信息实现电信诈骗的检测。

作者简介：郭开远（1994-），男，硕士研究生，主要研究方向：无线通信与移动计算
通信联系人：王文博（1965-），男，教授、博导，主要研究方向：无线移动通信. E-mail: wbwang@bupt.edu.cn

1 检测模型框架和数据预处理

1.1 诈骗用户检测模型框架

本文的研究点是基于运营商提供的蜂窝网络实测数据构建电信诈骗的检测模型，分析电信诈骗用户的特征，找到电信诈骗用户，并针对电信诈骗用户推出限制措施。检测模型具体包括 5 个步骤，如图 1 所示：



图 1 电信诈骗检测模型步骤

Fig.1 Telecom fraud detection model steps

数据准备：对目标群体进行定义，设计模型的数据宽表。根据数据宽表准备数据。

数据处理：分析诈骗用户的行为，根据特征的重要性和相关性进行特征筛选，并对数据集的数据进行编码，样本抽样和划分。

模型构建：以诈骗用户数据为正样本，采用本文改进的机器学习分类算法，构建疑似诈骗用户识别模型。

模型评估：评估模型的质量，选择效果最好的算法作为模型的算法。

模型应用：部署模型，根据模型结果检测出诈骗用户，针对诈骗用户推出限制措施。

1.2 数据预处理

1.2.1 数据集介绍

本文采用的蜂窝网络数据采集于广东省内某市电信运营商。数据已经过脱敏处理，ODS（Operational Data Store，操作数据存储）中已经进行了一定程度上的转换。论文经过对多类型电信诈骗场景进行分析，并将可获取的数据进行归纳，对数据的所有字段信息进行整理，初步获得如下特征，具体特征和特征分类如下：

用户基础属性：用户编码，用户号码，入网时常，证件名称，付费方式，用户归属地等用户基础信息在内的 53 个特征。

套餐信息：套餐信息指的是用户办理的套餐信息，包括套餐价格，套餐类型，套餐名称共 3 个特征。

终端信息：基于用户终端型号的信息，包括终端型号，终端类别，终端品牌等共 9 个特征。

上网行为：采集自用户的流量使用情况，包括日均流量，流量活跃天数占比共计 6 个特征。

通话行为：包括日均主叫次数，日均主叫时长，主叫时长 5 秒、5 秒-15 秒时间段占所有主叫次数的占比，各个时间段的主叫次数占比等基于用户通话行为的特征共 33 个。

短信行为：短信对象数，短信上下行条数共记 5 个特征。

基站数据：月均主叫基站数量，月均被叫归属地市数量，欺诈高发区域拨打次数占比等共计

26 个特征。

1.2.2 特征选择

特征选择的目的是筛选入模指标，达到降低训练复杂度的目的。原始数据中包含 134 个样本特征，对其进行特征选择，主要依据如下。

（1）特征的重要性较低

缺失值占比：属性缺失值超过 50%作为无效特征剔除。

分类型特征中类别占比：计算分类型变量中类别值的个数占总数的比例，如果比例超过 50%，则将该特征视作不重要特征，将该特征剔除。

分类型特征中类别个数占比：计算分类型变量中每个类别的个数占总数的比例，如果单个类别占比大于 90%，则将该特征视作不重要特征，将该特征剔除。

数值型特征的变异系数小于指定标准：变异系数是度量数值型特征取值离散程度的指标，如果数值型特征的变异系数大于 15%，则将该特征剔除。

根据以上依据，剔除的特征如表 1 所示：

表 1 剔除特征列表 1

Tab.1 List of Excluded feature 1

特征名	特征类型	特征剔除原因
证件名称	集合	单个类别过大
用户类型	集合	单个类别过大
新增用户标签	标志	单个类别过大
是否行业应用	标志	单个类别过大
是否黑名单	集合	单个类别过大
是否 4G 开卡	标志	单个类别过大
终端价格	集合	缺失值过多
终端型号	集合	所占类别比例过大
……	……	……

（2）特征变量与标签变量是否无关：

对于分类问题，与标签独立的特征为无关特征，特征选择的目的是去除无关特征。

具体方法与指标类型相关：对于分类型指标：进行卡方检验，卡方检验是统计学中评价两个事件是否独立的常用方法，通过卡方检验可以获得变量独立的概率 P，P 值越小说明该特征变量越重要，本文将 P 值小于 0.05 的特征视为重要的特征；对于数值型指标：进行 F 检验，F 检验是一类建立在 F 分布基础上的假设检验方法，分析目标变量不同取值下指标的均值是否有显著差异，有显著性差异的变量相关性强，通过 F 检验获得变量独立的概率 P，若 P 值小于 0.05，说明特征与标签相关，是重要的特征。

剔除的特征如表 2 所示：

表 2 剔除特征列表 2

Tab.2 List of Excluded feature 2

特征名	特征类型	特征是否重要	P 取值
是否融合套餐	标志	一般重要	0.082
客户拥有宽带产品数	连续	不重要	0.127
客户拥有 IPTV 产品数	连续	不重要	0.41
是否翼支付用户	标志	不重要	0.444
……	……	……	……

经过上述步骤，本文采用的数据集包含 63 个样本特征。

1.2.3 编码，样本抽样和划分

（1） 独热编码（One-Hot Encoding）

分类器存在不好解决的离散特征，不能处理字符串等问题，需要进行编码。本文采取独热编码的方式解决这些问题。

独热编码又称为一位有效编码，主要使用 N 位状态寄存器来编码 N 个状态，每个状态由其自己的寄存器位分隔，并且任何时候只有一个位有效。离散特征的值延伸到欧式空间，并且离散特征的特征值对应于欧式空间的某个点[1]。如自然状态码为：000,001,010,011,100,101，独热编码为：000001,000010,000100,001000,010000,100000。每个特征会变得互斥，让特征之间的距离计算更加合理。

（2） 样本抽样

原始样本中正样本的数量约为 69.9 万，负样本的数量约为 1.8 万，正负样本的比例为 1:37.8。为了提高建模效率，要对原始样本进行抽样，控制建模样本量；同时为了提高模型准确率，一般控制建模正负样本比例在 1:5 左右。该模型分别对正负样本进行随机抽样，本文所使用的模型分别对正负样本进行随机抽样，选择 6.7 万样本入模，其中正负样本比例约为 1:5。

（3） 数据集划分

为了保证模型的泛化能力，需要将数据集划分为训练集与测试集，本文将数据集随机划分，其中 70%作为训练集，30%作为测试集。

2 CNNcombine 算法

论文采用的数据集包含有一定数量的正样本。因此整个检测模型基于机器学习中分类算法的理论工具进行设计。论文基于卷积神经网络对机器学习中分类模型进行改进。

2.1 基于一维卷积神经网络的分类算法

2.1.1 一维卷积神经网络

卷积神经网络是一类深度前馈神经网络，一般由卷积层和池化层相互交替构成[2]。具有

局部连接、权重共享以及子采样的特性。这些特性使得卷积神经网络具有一定程度上的平移、缩放和扭曲不变性，并且卷积神经网络需要的参数更少，因此有更高的训练效率。卷积神经网络更新权重的方式是利用反向传播算法。

一维卷积神经网络（1D-CNN ，1D Convolutional Neural Network）的原理和特点与二维或三维的卷积神经网络大致相同。关键的不同是，1D-CNN 的输入是一维向量，神经网络中的卷积核与特征图也为一维。卷积神经网络通常运用于图像处理、自然语言处理的领域和音频/语音处理，本文创新性地将蜂窝网络数据集作为卷积神经网络的输入，将每个用户的特征视作一个一维向量，并可以通过 1D-CNN 的输出层实现二分类功能。

2.1.2 神经网络结构设计

本文设计的 CNN 卷积神经网络由输入层，5 个隐藏层，以及输出层构成，如图 2 所示。

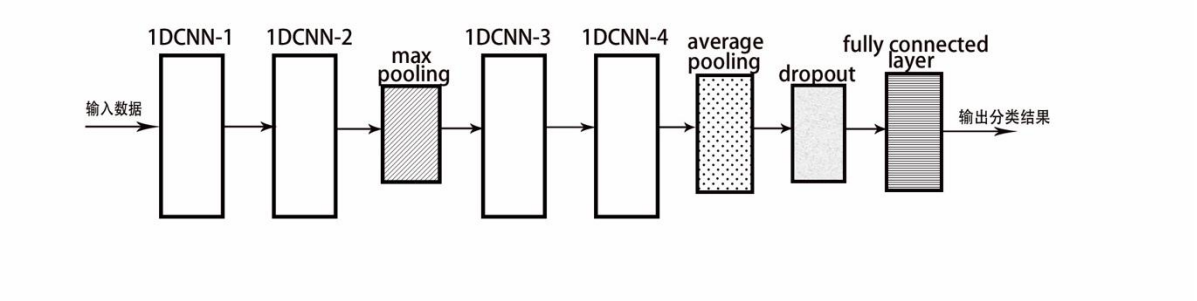


图 2 CNN 卷积神经网络结构

Fig.2 CNN convolutional neural network structure

第一层为输入层：根据前文预处理完成的数据，每条数据共计 63 个特征，因此需要将一个长度为 63 的向量传递到神经网络之中，因此这里需要将每个用户数据重新整形，构成一个 21 x 3 的矩阵作为神经网络的输入。

接下来的 2-7 层为隐藏层，是卷积神经网络中最重要的构成部分。

第一个 1D-CNN 层：定义高度为 2 的卷积核，过滤器将允许神经网络学习第一层中的单个特征，共定义 50 个过滤器，这允许在网络的第一层上训练 50 个不同的特征。输出是 20×50 神经元矩阵。输出矩阵的每列保持一个单个滤波器的权重。使用定义的内核大小并考虑输入矩阵的长度，每个过滤器将包含 20 个权重。

第二个 1D-CNN 层：单层的卷积只能获得浅层的抽象信息，为了增加神经网络的深度，再添加一个 1D-CNN 层。来自第一个 CNN 的结果将被送入到第二个 CNN 层。再次定义 50 个不同的过滤器，在此级别上进行训练。第二层遵循与第一层相同的逻辑，最终输出矩阵的大小为 19 x 50。

最大池化层（Max Pooling Layer）：通常在卷积层之后使用池化层，池化层的目地是降低输出的复杂性并防止数据过度拟合。最大池化层的原理是在邻域内取最大的特征值作为该区域的值，设置池化窗口大小为 3。这意味着该层的输出矩阵的大小仅为输入矩阵的三分之一。

第三和第四个 1D-CNN 层：再次添加两个卷积层以便学习更全局化的特征，卷积核大

小为 2，过滤器数量为 80。这两层的输出矩阵是 5×80 矩阵和 4x80 矩阵。共经过了四层卷积层，我们得到了包含全局特征的神经元，实现了特征提取。

全局平均池化层（Global Average Pooling Layer）：再添加一个池化层，以进一步避免过度拟合。全局平均池化层执行更极端类型的降维，每个特征检测器在该层上的神经网络中仅剩余一个权重。输出矩阵的大小为 1 x80 个神经元，为一维向量，可以视作经过神经网络特征提取后的特征向量，也为与其他分类算法结合提供了可能。

丢弃层（Dropout Layer）：添加一个丢弃层来解决过拟合和梯度消失的问题。丢弃层随机将网络中的部分神经元“丢弃”，设定 50％的神经元将权重为零。丢弃层取平均并减少神经元之间复杂的共适应关系的作用，达到解决过拟合的效果。该层的输出仍然是 1 x 80 的神经元矩阵。

最后用一个全连接（Full Connected）的形式，将输出层接入卷积神经网络。因为预测目标仅为两个类：电信诈骗用户和非电信诈骗用户，所以在最后一层将高度 80 的向量减少到高度为 2，通过该层的矩阵乘法完成。采取 Softmax 用作激活功能，它能够强制神经网络的所有 2 个输出总和为一。输出值将代表 2 个类中每个类的概率。

2.2 CNNcombine 算法原理

为了获得比 1D-CNN 和传统分类算法更优秀的分类精度，本文中我们提出一种分类模型 CNNcombine，该模型将 2.1 中设计的 1D-CNN 神经网络与传统的分类算法相结合。1D-CNN 的优势在于卷积层能够在输入的特征序列中提取高层特征，利用提取的特征来改进传统算法模型的准确性。

此外，单一的传统算法模型的在最佳参数下的预测结果即为模型的最好表现，不同模型采用的不同参数，很多情况下某一模型的预测结果不是最好[3]，但是模型对于样本集中的一些样本的预测效果可能会好于原先最佳的预测模型。即每个模型的预测结果不相同，存在着通过“互补”的方式以取得更好的预测准确性的可能，因此通过模型融合的方式，可以获得更优的模型。

本文将传统算法进行模型融合的思路是通过元分类器对多个分类器进行融合，次级分类器训练后，基于训练数据给出预测结果，元模型基于次级分类器的预测结果的输出再次进行训练。次级分类器可以采用多种不同的分类算法，即可以提高模型的泛化能力也将每个次级分类器执行的在不同数据的优点体现出来，因此模型融合后是异构型。将数据集均分为 4 个部分，其中三个作为训练集，一个作为测试集。AdaBoost[4]、随机森林[5]、GBDT[6]和 XGBoost[7]作为集成学习算法，单独算法已经在蜂窝网络数据集上的预测有比较好的预测效果，因此论文将这四种算法作为融合的基础。首先选择 AdaBoost、随机森林、GBDT 作为次级分类器，对训练集进行 K 折交叉训练，根据输出结果的标签作为数据集新的特征，构成新的数据集。之后选择 XGBoost 作为元分类器，根据新的数据集训练元分类器，最终输出综合模型的预测结果。

结合 1D-CNN 的实现方法是，用综合算法的分类器替代 1D-CNN 中最后的全联接层，

蜂窝网络数据集经过神经网络的进行特征提取，输入综合算法的特征向量长度为 80。

CNNcombine 算法的完整结构如图 3 所示：

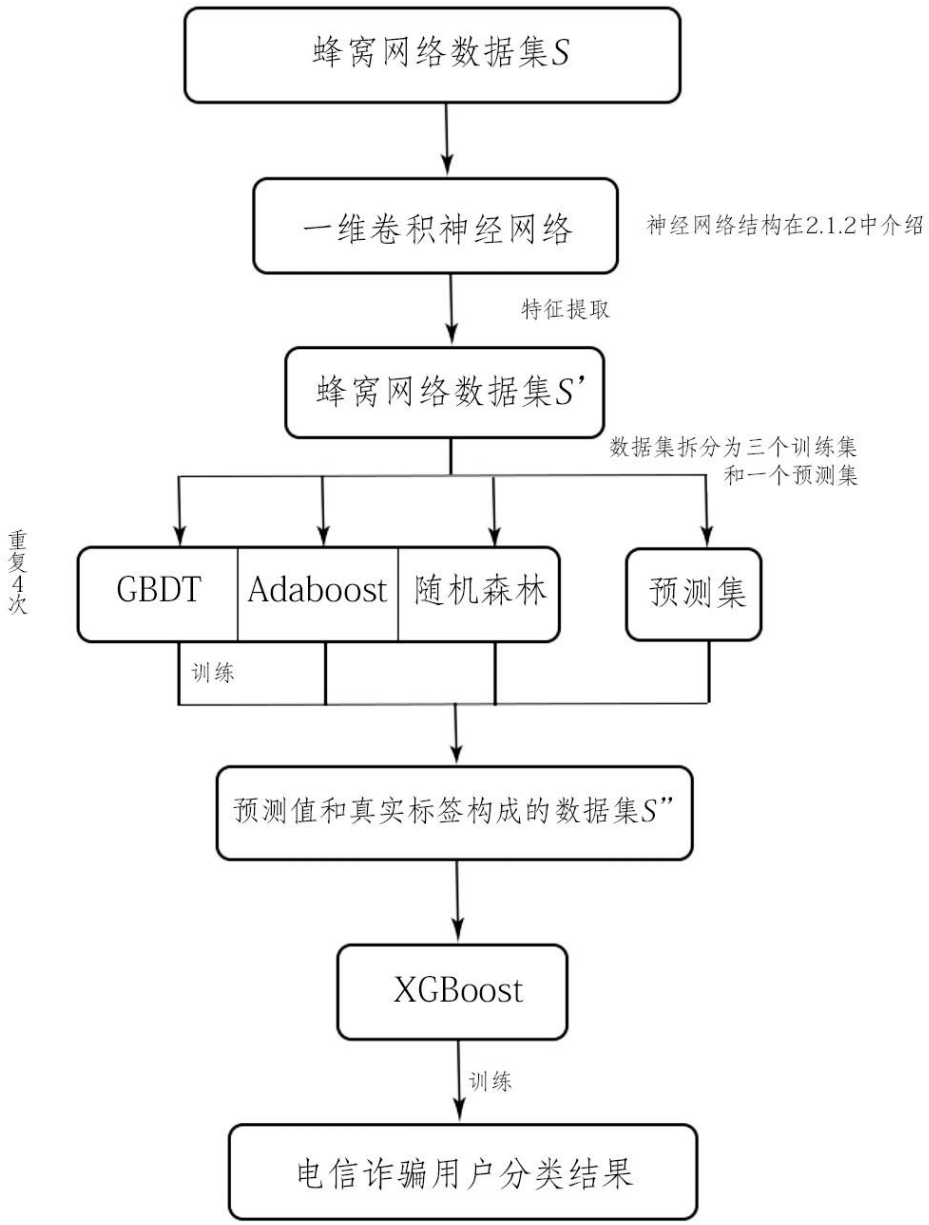


图 3　CNNcombine 算法结构

Fig.3　CNNcombine algorithm structure

3　模型评估

本节使用来自电信运营商的数据集来验证模型检测电信诈骗号码的有效性，设计实验，比较文中提出的算法和常用的传统分类算法在性能上的差异。

3.1　实验设计

将数据集随机划分，其中 70%作为训练集，30%作为测试集。使用该数据集在 Python

环境下进行分类算法训练，选取的算法有本文提出的 CNNcombine 分类模型和设计的 1D-CNN 分类模型，并以 XGBoost 和 AdaBoost 两种被广泛运用的传统分类方法作为对照组。

由于本文所采用的分类器的输出，输出结果不是 0 或 1，而是实数值，表示的是该用户属于电信诈骗用户概率，预测结果的值越大，说明样本属于正样本即电信诈骗用户的概率越大，反之值越小则说明为正常用户概率越大。根据需求的不同可以设置不同的阈值，通过分类算法的结果与阈值的比较判断样本是否属于正样本。使用 P-R 曲线 (Precision-Recall Curve) [8]以及平均精度 AP（Average precision）[8]作为实验的评估标准。

P-R 曲线中，P 表示精确率，R 表示召回率。以 P 作为横坐标，R 作为纵坐标，对于同一个模型，根据逐个样本作为阈值划分点的方法，可以得到一条曲线，这就是 P-R 曲线。假设一个模型的 P-R 曲线能完全包围另一个学习器，说明前者的性能更好。

AP 是用于概括 P-R 曲线的单个指标，它表示在每个阈值处实现的精确率的加权平均值，从先前阈值的召回率的增加作为权重。AP 的数学表达式如下：

$$AP=\sum_n (R_n - R_{n-1})\ P_n \qquad (1)$$

其中 R_n 和 P_n 分别表示输出结果在第 n 个阈值处的精确率和召回率。AP 的数值越大，说明分类的结果越好。

3.2 性能评估结果

对比四种不同的分类模型的分类结果，绘制的 P-R 曲线由图 4，图 5 所示，其中图 5 放大了 P-R 曲线的右上角部分。

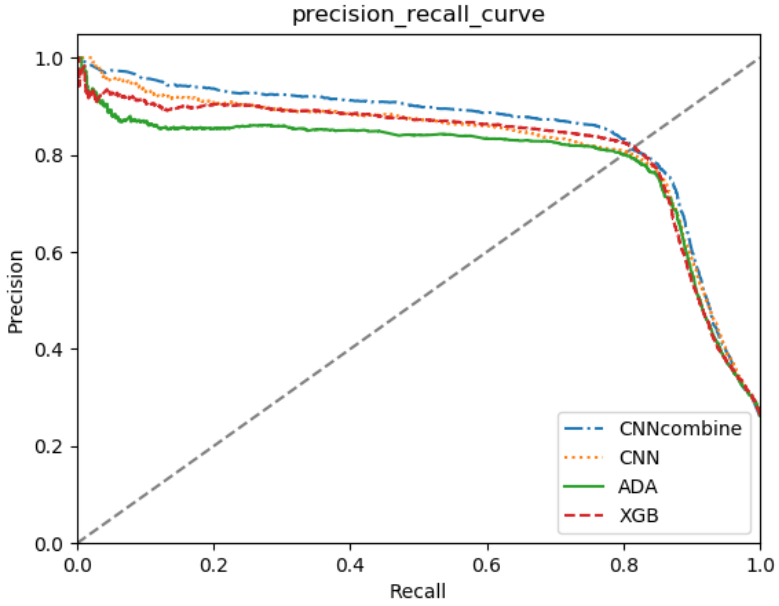


图 4 P-R 曲线对比

Fig.4 CNNcombine algorithm structure

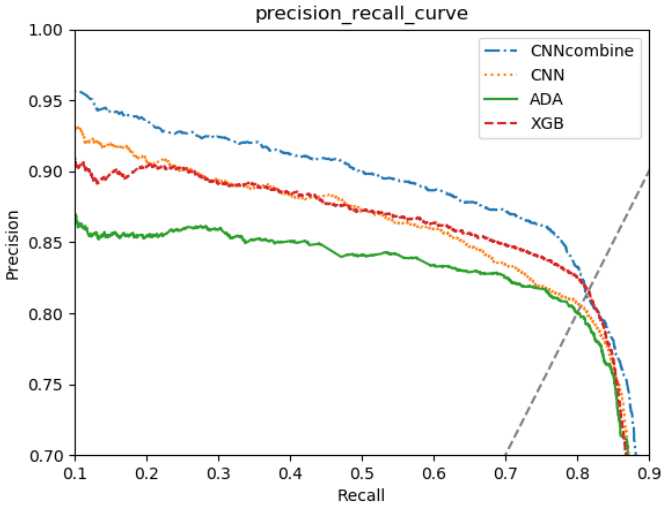


图 5　P-R 曲线对比（右上角放大）

Fig.5 P-R curve comparison (upper right corner)

图 6 的柱状图作为更直观的对比，采用的评估的标准为平均精度 AP：

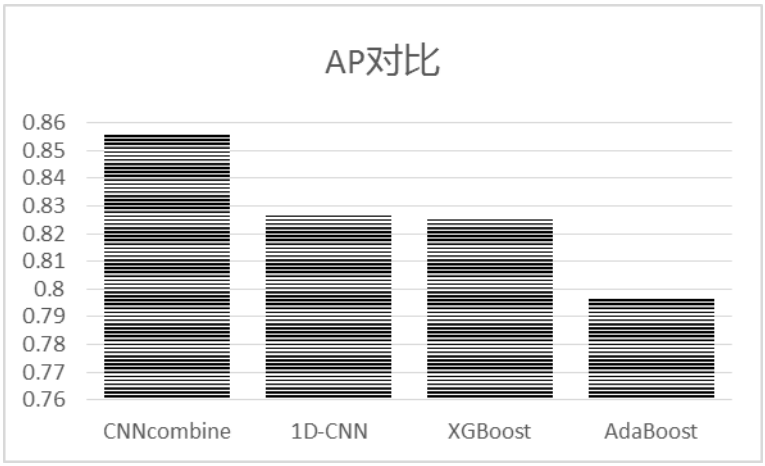


图 6　平均精度 AP 对比

Fig.6　Average Precision comparison

综合以上模型结果评估可以证明，改进算法 CNNcombine 在检测电信诈骗用户上的分类性能优于其他的分类模型，另一方面将本文设计的 1D-CNN 神经网络直接用于应对现实中检测电信诈骗用户的问题时，效果对比 XGBoost 这样传统的分类算法没有明显的优势。实验评估证明，CNNcombine 算法的电信诈骗检测模型能够更准确的从众多用户中检测到诈骗用户。

电信诈骗检测模型的输出结果是该用户为诈骗用户的概率，在实际运用中需要设定分类阈值来确定预测结果，即预测概率大于该阈值的分类为诈骗用户，预测概率小于该阈值的则分类为正常用户。通常阈值的确定方法是选择精确率与召回率相同时对应的阈值，此时能同时保持较高的精确率和召回率，精确率和召回率相同时测试集检测模型的分类阈值为:0.41777，模型的精确率和召回率为 0.8155，在测试数据集中 5242 个诈骗用户中能够正确检测到 4275 位诈骗用户，对测试集全体 20133 个用户的分类准确率为 90.394%。

4 结论

本文研究的内容是设计与实现电信诈骗用户的检测模型，并提出一种将一维卷积神经网络与传统分类算法相结合的 CNNcombine 算法。本文基于蜂窝网络实测数据的实验，验证了 CNNcombine 算法对比 XGBoost、Adaboost 这些传统算法有更好的预测结果，精确率与召回率都有明显的提升，平均精度比 XGBoost 高 3%，比 Adaboost 高 6%，证明本文提出的模型在解决实际问题中的有效性和可行性。

[参考文献] (References)

[1] Xlinx. HDL Synthesis for FPGAs Design Guide[M]. XACT, 1995: 3-13
[2] Takahashi N, Nishi T, Hara H. Analysis of signal propagation in 1-D CNNs with the antisymmetric template[A].12th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA 2010)[C].Berkeley:IEEE.2010 XIAO J Z, LEI B, WANG C Q. Reclamation on building waste produced from Wenchuan Earthquake[A]. XIAO J Z. The first discussion forum of study on RAC[C]. Shanghai: Tongji University Press, 2008. 64-65.
[3] Sill J, Takacs G, Mackey L, et al. Feature-Weighted Linear Stacking[J]. Computer Science, 2009.
[4] Yoav Freaud, Robert E. Schapire. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting[J].journal of computer and system sciences 55, 119-139
[5] L B. Random Forest [J]. Machine Learning, 2001, 45: 5-32.
[6] Friedman J H.Greedy function approximation:a gradient boosting machine[J]. Annals of statistics, 2001:1189-1232.
[7] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J].In Proceeding KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016: 785-794.
[8] 周志华.机器学习: Machine learning[M].北京:清华大学出版社, 2016,26-48