研究与开发

改进自训练模型在业务质差用户识别中的应用

余立[1]，李哲[1]，高飞[1]，袁向阳[1]，杨永[2]
（1. 中国移动通信有限公司研究院，北京 100053；
2. 中国移动通信集团公司，北京 100033）

摘　要：质差用户识别是降低用户投诉率、提升用户满意度的重要环节。针对当前电信网络系统中业务感知相关的大量结构化及非结构化数据难以有效标注、质差用户标签不完备、现有监督学习模型训练样本不均衡而导致质差识别率低的问题，采用改进自训练半监督学习模型，利用少量满意度低分和投诉用户作为质差用户标签对网络数据进行标注，并通过标签迁移对大量未标注数据进行训练识别质差用户。实验表明，相比于识别准确率高但是训练成本高的全监督学习和识别准确率低的无监督学习，半监督学习可以充分利用无标签样本数据进行有效训练，保证较低训练成本的同时显著提升质差用户识别准确率。

关键词：半监督学习；改进自训练模型；质差用户识别；无标签数据
中图分类号：TN915.41
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2021191

Application of improved self-training model in the identification of users with poor service quality

YU Li[1], LI Zhe[1], GAO Fei[1], YUAN Xiangyang[1], YANG Yong[2]
1. China Mobile Research Institute, Beijing 100053, China
2. China Mobile Communications Corporation, Beijing 100033, China

Abstract: Poor quality user identification is an important method to reduce the complaint rate and increase satisfaction. It is difficult to effectively label a large amount of structured and unstructured data related to business perception in current telecommunications network systems, poor quality user labels are not complete, and the existing supervised learning model training samples are unbalanced, resulting in a low quality recognition rate. An improved self-training semi-supervised learning model was adopted, a small number of low-satisfaction and complaint users as poor quality user labels was used to label network data, and label migration was used to train a large amount of unlabeled data to identify poor quality users. Experiments show that compared to fully supervised learning with high recognition model accuracy but high training cost and unsupervised learning with low recognition model accuracy, semi-supervised learning can make full use of unlabeled sample data for effective training, ensuring lower training costs and the recognition accuracy of poor-quality users is significantly improved.

Key words: semi-supervised learning, improved self-training model, poor quality user identification, unlabeled data

收稿日期：2021−05−01；修回日期：2021−06−20

1　引言

随着移动互联网发展，我国移动互联网用户突破 13 亿户，占全球网民规模的 32.17%[1]，随着新型技术（如 5G、云计算等）的发展，用户对上网速度、稳定性等要求越来越高。

质差用户指在使用移动通信网络服务时，由于网络质量问题或其他因素对服务体验不满的用户。网络质量问题导致的质差用户，对网络服务的满意度会降低，且可能存在投诉、转网等行为。

质差用户群体流失概率较高，他们是各大网络运营商重点关注与关怀对象。传统质差用户识别通过数据采集系统对用户上网过程中产生的行为单据 XDR（X data record）进行分析，即可过滤潜在的质差用户。但各用户感知无法统一，不满意原因及不满意的业务也并不一致，传统分析方法可识别的投诉用户比例较低，无法满足现网投诉处理要求。故通过将已存在的满意度低或投诉行为的质差用户与 XDR 进行关联标注后，利用机器学习算法，实现对质差用户进行分类识别与预测。

通过现网收集的 XDR 数据中存在以下问题。

- 不同省份网络基础设备存在一定差异，数据中特征字段并不完全相同，且部分字段填充率较低，无法直接利用。
- 数据进行标签化标注时，不同省份字段计算方法可能存在差异，且数据量巨大，导致标注成本高昂。
- 已投诉用户单据中投诉原因众多，部分原因来自于非网络问题，存在大量对抗样本，导致样本本身含有较大噪声，训练时会影响模型性能。

2　半监督学习

有监督学习需大量的有标签数据训练模型，在质差用户识别模型中，一条有标签数据包含两部分：用户 XDR 数据和是否为质差用户。前部分数据通过数据采集系统获得，后部分标签信息需丰富的专家知识，往往判定成本较高，造成整体训练成本的增加[2]。

现网每日会产生海量的用户 XDR 数据，通过标注再进行训练，模型时效性较差，无法准确描述现网实时运行状况。半监督学习与有监督学习相比可以利用现网实时海量无标签数据，效率较高；与无监督学习相比可以保证模型准确率。

质差用户识别为分类问题，已知的半监督分类问题主要分为 5 类，具体优劣势见表 1。

- 基于图的半监督模型将标签数据和未标签数据构建为图，图中节点为数据点，边为节点权重，通过寻找图的最小分割，然后计算反向传播权重，其可应用于图片、中文文本、数据分类等各类场景，但是当新样本加入时，需要重新训练得到图模型，计算开销较大[3-4]。
- 基于分歧的半监督模型通过选择差异化基

表 1　半监督分类方法优劣势

方法	参考文献	优势	劣势
基于图的半监督	[3-4]	应用场景广泛、拓展性强	新样本加入重新训练
基于分歧的半监督	[5-6]	多分类器合作，提高准确率	不同分类器性能要求高
半监督支持向量机	[7]	数据样本需求较少	可解释性差、准确率受限
协同训练	[8-9]	提高模型预测精度	数据预处理要求较高
自训练	[10-12]	运行效率快，精度较高	错误累计会导致性能下降

模型，进行组合降低“错误”分类样本对模型的不良影响，提升模型预测准确率，但是其对基模型选择设定要求较高，并且运算效率也较低[5-6]。

- 半监督支持向量机是将支持向量机应用到半监督模型中，将样本空间映射到高维空间，并选择合适平面将样本集划分，但是模型受参数影响，最终模型准确率较低[7]。
- 协同训练（co-training）用有标签样本的两个视图分别训练两个弱分类器，再利用分类器对未标注样本预测中高置信度样本训练另一个分类器；即用一个视图中获得的知识来训练另一个视图。缺点是对样本要求高，要求具有两个充分冗余且满足条件独立性的视图，实际情况下较难满足[8-9]。
- 自训练（self-training）需要一个基分类器和少量样本数据可以实现，核心思想是先学习有标签数据，然后计算无标签样本置信度，并将置信度高的样本加入训练集，缺点是如果无标签样本预测错误，则随着训练的深入，会造成错误的累计[10-12]。

基于对以上半监督方法的研究，本文选取一种改进自训练模型，通过设置基模型参数以及较高的置信度阈值，引入多个基模型学习器，降低传统自训练中出现的误差累计现象，提高模型训练精度。

3 改进自训练应用

3.1 改进自训练模型

自训练模型是一种增量模型，首先建立基分类器模型，通过有标签数据进行训练，然后利用训练好的基模型不断预测数据集中无标签数据，从中选择置信度高样本，将其加入有标签数据中进行基模型循环训练。在满足设定停止迭代条件后，得到具有最高分类精度和最强的泛化性能的最终分类器。模型在迭代过程中不可避免会产生误分样本，基模型学习误分样本会产生错误累计，最终影响模型效果。为降低错误累计，本文做出以下改进：设置模型性质不同、性能相同的 3 种基模型，分别进行预测后，通过投票初步选定伪标签样本，随后计算其置信度，将置信度高的伪样本加入模型训练集中进行循环迭代。改进自训练模型示意图如图 1 所示。

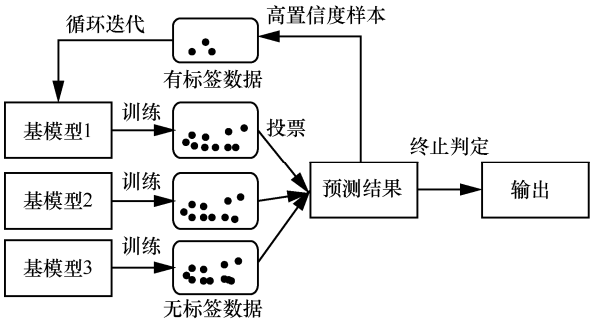


图 1 改进自训练模型示意图

其基本流程如下：

（1）根据有标签数据集训练 3 种基模型；

（2）利用训练得到的基模型预测无标签数据；

（3）选择置信度高的样本，将其加入有标签数据集；

（4）循环训练模型；

（5）判断是否满足迭代条件，重复（1）～（3）。

改进自训练模型算法见算法 1。

算法 1

输入 有标记样本集：$D_l=\{(<x_1^1,x_1^2,\cdots,>,y_1),\cdots,(<x_l^1,x_l^2,\cdots,>,y_l)\}$；

无标记样本集：$D_u=\{(<x_{l+1}^1,x_{l+1}^2,\cdots,>),\cdots,(<x_{l+u}^1,x_{l+u}^2,\cdots,>)\}$；

每 i 轮基学习器为 K_{1i}、K_{2i}、K_{3i}；

每 i 轮预测得到样例数为 p_i；

流程

（1）初始化设置 K_{10}、K_{20}、K_{30}；

（2）$i=1$；

（3）利用 K_{10}、K_{20}、K_{30} 拟合 D_l 得到 K_{11}、K_{21}、K_{31}；

（4）利用 K_{11}、K_{21}、K_{31} 训练 D_u，得到 p_i 例样本不同分类情况下置信度；

（5）进行选择，将 p_i 例预测样本加入 D_l；

（6）利用新样本集 D_l 训练，得到 K_{12}、K_{22}；

（7）循环（4）～（6），直到满足迭代终止条件。

3.2　基模型选取

质差用户识别是一种分类问题，最终评价标签为质差用户和非质差用户两种。当前主流机器学习分类模型有贝叶斯分类器（NB）、Logistic 模型、支持向量机、树模型（如随机森林（RF）和极限梯度提升（XGBoost）等[13]）。本模型选择朴素贝叶斯分类器、XGBoost、随机森林为基模型进行训练。

（1）朴素贝叶斯分类器

该模型描述如下：设训练集中包含 m 个类 $H=(H_1,H_2,\cdots,H_m)$，n 个条件属性 $X=(X_1,X_2,\cdots,X_n)$，并且假设所有条件属性 X 为类变量 H 的子节点，并相互独立，则当待分类样本 $x=(x_1,x_2,\cdots,x_n)$ 分配到类 H_m 时，根据贝叶斯定理可得：

$$p(H_m|X)=\frac{p(X|H_m)\,p(H_m)}{p(X)} \tag{1}$$

由于在本监督学习中需要使用大量的无标签数据进行模型训练，式（1）修改为：

$$p(H_m|X)=\max P(H_m)(n+l)\prod_{m=1}^{n}p(X|H_m)(n+l) \tag{2}$$

其中，$(n+l)$ 表示迭代后有标签样本集与增加标记后的无标签样本集的合集，该合集增加了无标签数据中预测得到的高置信度数据。

（2）极限梯度提升和随机森林

XGBoost 和 RF 都是基于树模型的集成模型，但是两者有所区别。XGBoost 为并行化 Boosting 处理，RF 为串行化 Bagging 处理[14]。

给定数据集 $D=(X_i,y_i)$，输入 X_i 并通过线性叠加模式预测 y_i。并设学习使用 k 棵树，模型如式（3）、式（4）所示。

$$\widehat{y_i}=\varnothing(X_i)=\sum_{k=1}^{K}f_k(X_i),f_k\in F \tag{3}$$

$$F=\{f(X)=\omega_q(X)\}(q:R^m\to T,\omega\in R^T) \tag{4}$$

其中，$f(X)$ 代表回归树，F 代表回归集合，$q(X)$ 表示将 X 分到了某个叶子节点上，T 为叶子节点的数量，ω 为叶子节点分数，$\omega_q(X)$ 代表 $f(X)$ 对样本的预测。

通过二阶泰勒展开式和正则项调整得到目标函数如式（5）所示。

$$\begin{aligned}\mathrm{Obj}^t\simeq&\sum_{i=1}^{n}l\left(y_i,\widehat{y_i}^{(t-1)}\right)+g_if_t(x_i)+\\&\frac{1}{2}h_if_t^2(x_i)]+\Omega(f_t)+\mathrm{constant}\end{aligned} \tag{5}$$

4　实验与分析

4.1　数据预处理

（1）数据收集

本次仿真实验使用数据采集系统中的正常 XDR 数据和现网投诉 XDR 数据。其中包含 199 998 条、46 项字段的正常 XDR 数据以及 3 355 条、125 项字段的投诉 XDR 数据。

（2）字段选取

正常 XDR 数据的 46 项字段中，包含较多非结构化离散字段（如小区 ID、所属城市、IPV 类型等），并且部分字段缺失值比重较大，通过处理最终得到连续型字段 15 项。

（3）标准化处理

部分机器学习模型需要数据处于同一量纲，所以进行数值量纲转化、标准化处理。处理后数据变化到均值为 0、方差为 1 范围内。

（4）样本均衡

因为质差用户识别为二分类问题，所以需要保证原始训练数据样本集分布相同。使用随机采样方法，最终案例数据集组成见表 2。

表 2　数据集组成

数据类型	数据条数/条	属性值/列	标签/列
质差用户	3 000	15	1
非质差用户	3 000	15	1

（5）关键参数

TCP 建链成功到第一条事务请求的时延（tcp_ack_srv_dur）：在终端和服务器完成 TCP 建链请求后，到终端发出业务请求前的时间间隔。

第一个 HTTP 响应包时延（fisrt_http_ response_ time）：在业务请求过程中，第一次业务请求发出后到接收第一次业务请求响应的时间间隔。

TCP 建链确认时延（fisrt_http_response_ time）：在 TCP 建链过程中，第二次握手 SYNACK 报文发出后到收到第三次握手 ACK 报文的时间间隔。

4.2　分类器评价标准

对于每个待检测的用户数据，分类器最终可能产生 4 种不同的结果，本实验中对不同情况解释如下。

- TP（true positive）：质差用户，且模型预测结果为质差用户。
- TN（true negative）：非质差用户，且模型预测结果为非质差用户。
- FP（false positive）：非质差用户，但模型预测结果为质差用户。
- FN（false negative）：质差用户，但模型预测结果为非质差用户。

基于以上 4 种情况，引入精准度、F1 值和 AUC3 项指标进行评判。精准度和 F1 主要判断分类器预测结果的准确性，AUC 主要判断分类器对质差用户区分能力的强弱。

精准度即精确率，在本实验中表示正确判断为质差用户的样本占全部质差样本的比例：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

F1 值是由 Precision 和 Recall 的调和平均数，在本实验中表示在保持一定精确率同时，尽可能保证所有质差用户可以被模型识别即保证召回率，两者相互平衡。

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

AUC 值是 ROC 曲线下方的面积。ROC 曲线绘制的横坐标是 FPR，而纵坐标是 TPR。当无法直接衡量学习性能时，AUC 值越大，表明模型效果越好。

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{9}$$

4.3　实验结果与对比分析

针对实验数据，分别使用全监督方法、半监督方法、无监督方法进行拟合。其中全监督方法选用 XGBoost 模型，半监督方法使用本文提出的改进自训练模型，无监督方法选用图传播 label spreading 模型。结果对比见表 3。

表 3　3 种模型运行结果

模型	精准度	AUC	F1
全监督模型	0.970 6	0.970 6	0.969 9
半监督模型	0.939 4	0.938 9	0.936 5
无监督模型	0.740 5	—	—

对比 3 类模型精准度可得到如下结论：全监督模型效果最好，各项评价指标数值最高；无监督模型效果最差，因为在模型训练过程中不可避免会学习到数据中噪声，影响模型评价指标；而半监督模型介于两者之间，可以充分利用大量无标签数据，此外还可以保证较高精准度。

为了进一步验证半监督模型优越性，将以上 3 类模型进行对比。其中在半监督和无监督模型中，横轴设置为样本标签缺失值比率。在全监督模型中，横轴设置为训练集划分比率。不同缺失值比率下模型精准度变化如图 2 所示。

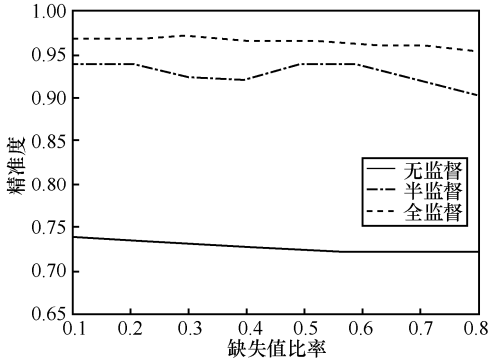


图 2　3 类模型精准度对比

通过图 2 可知，随着缺失值比率增加，3 类模型精准度都在下降，半监督模型仍然处于一定精准度变化区间内，可以满足模型识别精准度要求。

在半监督改进自训练模型中，使用迭代训练对无标签数据进行标注，当缺失值比率相比于上次训练变化幅度低于 0.1%时，模型迭代停止。具体数值和曲线变化如图 3 和表 4 所示。

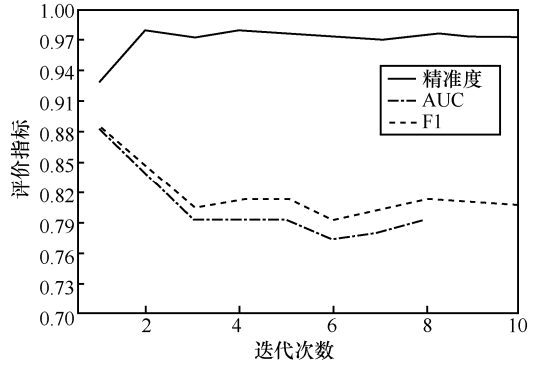


图 3　半监督模型评价指标变化趋势

表 4　半监督模型变化趋势

轮次	缺失值比率	精准度	AUC	F1
1	0.900 0	0.918 4	0.884 6	0.878 1
2	0.367 9	0.975 5	0.816 8	0.825 4
3	0.267 9	0.976 6	0.810 7	0.819 3
4	0.157 6	0.976 8	0.812 5	0.818 8
5	0.091 9	0.976 7	0.801 7	0.810 5
6	0.066 9	0.968 9	0.803	0.812 8
7	0.056 9	0.977 5	0.796 8	0.806 5
8	0.030 5	0.978 1	0.799 1	0.809 0
9	0.026 7	0.974 3	0.792 8	0.805 4
10	0.026 0	0.978 2	0.796 8	0.806 1

如表 4 所示，设置默认参数和初始样本缺失值比率（Ratio）后，模型开始训练。通过 10 轮迭代计算后，Ratio 变化幅度 0.07%符合迭代终止条件，迭代停止。观察表 4 中数据可知，缺失值比率列数值下降明显，精准度、AUC、F1 3 项评价指标有一定波动，且浮动下降。这是因为模型在自训练过程中不可避免会学习到样本集中的噪声，最终模型性能受到一定影响。为进一步提升模型识别精准度，在之后的模型训练过程中，需改进基模型选择设计方案，并通过提高阈值、增加样本预测可靠性水平等方法，降低训练过程的误分类样本噪声。

5　结束语

本文针对质差用户识别问题，设计一种改进自训练的半监督模型，采用无标签样本占 90%的训练集时，最终模型精准度维持在 90%左右。相比于全监督模型和无监督模型，该模型在保证一定性能指标前提下，能够充分利用无标签样本数据，在现网应用中可有效降低数据标注成本，同时避免了人为主观因素对于质差规则设定的影响，可以有效实现质差用户识别。未来的工作重点为进一步提高该模型性能，降低在循环迭代中噪声对于模型性能的影响。

参考文献：

[1] 麻瓯勃, 刘雪娇, 唐旭栋, 等. 基于半监督学习的恶意 URL 检测方法[J]. 计算机系统应用, 2020, 29(11): 11-20.
MA O B, LIU X J, TANG X D, et al. Malicious URL detection based on semi-supervised learning[J]. Computer Systems & Applications, 2020, 29(11): 11-20.

[2] 欧阳晔, 杨爱东, 孟凡语. 一种博弈论辅助的机器学习算法检测用户流失行为[J]. 电信科学, 2020, 36(6): 79-89.
OUYANG Y, YANG A D, MENG F Y. A game theory-assisted machine learning methodology for subscriber churn behaviors detection[J]. Telecommunications Science, 2020, 36(6): 79-89.

[3] 张俊丽, 常艳丽, 师文. 标签传播算法理论及其应用研究综述[J]. 计算机应用研究, 2013, 30(1): 21-25.
ZHANG J L, CHANG Y L, SHI W. Overview on label propagation algorithm and applications[J]. Application Research of Computers, 2013, 30(1): 21-25.

[4] 彭杰, 龚晓峰, 李剑. LPA-SKFST 半监督特征提取方法[J]. 计算机应用研究, 2021, 38(6): 1657-1661.
PENG J, GONG X F, LI J. Semi-supervised feature extraction method based on LPA-SKFST[J]. Application Research of Computers, 2021, 38(6): 1657-1661.

[5] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11): 1871-1878.
ZHOU Z H. Disagreement-based Semi-supervised Learning[J]. Acta Automatica Sinica, 2013, 39(11): 1871-1878.

[6] ZHOU Z H, LI M. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[7] HUANG H J, QIAN L, WANG Y J. A SVM-based technique to detect phishing URLs[J]. Information Technology Journal, 2012, 11(7): 921-925.

[8] NIGAM K, GHANI R. Analyzing the effectiveness and applicability of co-training[C]// Proceedings of the ninth international conference on Information and knowledge management. [S.l.:s.n.], 2000: 86-93.

[9] ZHOU Z H. Disagreement-based semi-supervised Learning[J]. Acta Automatica Sinica, 2013, 39(11): 1871.

[10] MCCLOSKY D, CHARNIAK E, JOHNSON M. Effective self-training for parsing[C]// Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Piscataway: IEEE Press, 2006.

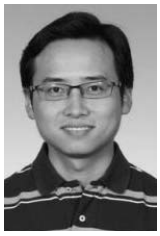
[11] ROSENBERG C, HEBERT M, SCHNEIDERMAN H. Semi-supervised self-training of object detection models[C]//Proceedings of 2005 Seventh IEEE Workshops on Applications of Computer Vision. Piscataway: IEEE Press, 2005: 29-36.

[12] 黎隽男. 半监督自训练方法的研究[D]. 重庆: 重庆师范大学, 2018.
LI J N. Research on semi-supervised self-training method[D]. Chongqing: Chongqing Normal University, 2018.

[13] 董立岩, 隋鹏, 孙鹏, 等. 基于半监督学习的朴素贝叶斯分类新算法[J]. 吉林大学学报(工学版), 2016, 46(3): 884-889.
DONG L Y, SUI P, SUN P, et al. Novel naive Bayes classification algorithm based on semi-supervised learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(3): 884-889.

[14] 陈颖, 杨欣, 孙道贺. 基于 GA-XGBoost 模型的大学生科研能力预测问题研究[J]. 数学的实践与认识, 2021, 51(6): 318-328.
CHEN Y, YANG X, SUN D H. Research on college students' scientific research ability prediction based on GA-XGBoost model[J]. Mathematics in Practice and Theory, 2021, 51(6): 318-328.

[作者简介]



余立（1981− ），男，中国移动通信有限公司研究院人工智能与智慧运营中心副总经理、高级工程师，主要研究方向为前沿移动通信技术、网络智能化、大数据和 IT 技术。

李哲（1992− ），男，中国移动通信有限公司研究院研究员，主要研究方向为 5G 核心网、人工智能、电信大数据、深度报文解析。

高飞（1978− ），男，中国移动通信有限公司研究院研究员，主要研究方向为人工智能、网络大数据分析、数据治理。

袁向阳（1978− ），男，中国移动通信有限公司研究院人工智能与智慧运营中心副总经理，主要研究方向为人工智能、网络智能化、大数据和 IT 技术。

杨永（1972−），男，中国移动通信集团公司网络事业部服务保障室经理，主要研究方向为无线网络质量、业务指标规划分析。