计算机科学
COMPUTER SCIENCE

非完美多分类标签体系下的领域短文本分类方法研究

梁浩玮，王石，曹存根

引用本文

梁浩玮，王石，曹存根.非完美多分类标签体系下的领域短文本分类方法研究[J]．计算机科学, 2023, 50(1): 185-193.

LIANG Haowei, WANG Shi, CAO Cungen. Study on Short Text Classification with Imperfect Labels[J]. Computer Science, 2023, 50(1): 185-193.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100278

非完美多分类标签体系下的领域短文本分类方法研究

梁浩玮　王　石　曹存根
中国科学院计算技术研究所　北京 100190
(liang199611@outlook.com)

摘　要　近年来,短文本分类技术获得了广泛的研究。但在实际应用中,随着文本数据的积累,人们经常会遇到分类体系问题及其引起的数据分类标注问题,原因在于分类标签体系通常具有动态性,以及体系中的分类标签具有不易区分性。为此,文中结合分类标签数量众多的某省电信投诉工单分析业务进行了具体分析,并提出了一种非完美多分类标签体系的概念模型。在此基础上,针对数据集中的分类标注冲突与遗漏,提出了一种基于高质量种子训练集的检测和半自动修复方法,用于修复分类体系动态性和人工标注错误导致的标注冲突和遗漏,经过6个月的线上运行,在过滤掉10%的分类置信度过低的投诉工单后,基于BERT的分类模型的F1值可达0.9。

关键词:非完美多分类标签体系;细粒度短文本分类;分类标注;数据清洗
中图法分类号　TP391

Study on Short Text Classification with Imperfect Labels

LIANG Haowei, WANG Shi and CAO Cungen
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract　Short text classification techniques have been widely studied. When these techniques are applied to domain short text forproduction, as textual data accumulates, people often encounter problems mainly in two aspects: the imperfect labels and mistakenly-labeled training dataset. First, the class label set is generally dynamic in nature. Second, when domain annotators label textual data, it is hard to distinguish some fine-grained class label from others. For the above problems, this paper analyzes the shortcomings of an actual and complex telecom domain label set with numerous classes in depth and proposes a conceptual model for the imperfect multi-classification label system. Based on the conceptual model, for repairing the conflicts and omissions in a labeled dataset, we introduce a semi-automatic method for detecting these problems iteratively with the help of a seed dataset. After repairing the conflicts and omissions caused by a dynamic label set and mistakes of annotators, after about six months of iteration, the F1-score of the BERT-based classification model is above 0.9 after filtering out 10% tickets with low classification confidence.

Keywords　Imperfect multi-classification label system, Fine-grained short text classification, Class labeling, Data cleaning

1　引言

领域文本分类是各行各业的基本任务。领域文本的形式很多,有长文本如图书、新闻、论文等;短文本包括投诉工单、市民反馈、销售工单等,本文的工作面向短文本。对领域文本进行分类分析是进行产品改进、风险预测和服务质检等工作的重要基础。

文本分类技术历经了几十年的广泛研究,在各行各业中得到了大量的实际应用。文本分类的方法有很多,近年来基于深度预训练语言模型的方法在性能上表现出众[1]。但是,在实际应用短文本分类算法的过程中,随着文本数据的积累,我们经常会遇到两个问题,即分类体系问题及其引起的数据分类标注问题,具体原因为:

(1)分类标签体系通常具有动态性。领域专家建设合适的分类标签体系,领域人员使用其进行业务分类和处理。但是,随着业务的发展以及认识的加深,领域专家在不断修改和完善分类标签体系。旧的分类标签体系不及时和不细致的缺点使得体系在更新之后与旧体系之下标注的数据相比呈现出标注冲突或遗漏;分类标签体系不简要和不全面的缺点导致标注人员在分类标注过程中感到非常疑惑,因此可能导致不同标注人员的分类标注结果出现冲突。

(2)分类标签体系中的分类标签具有不易区分性。人工

到稿日期:2021-11-29　返修日期:2022-09-01
基金项目:科技部重点研发计划课题:开放式智能化中医传承信息管理和挖掘平台的研制(2017YFC1700302)
This work was supported by the Development of an Open and Intelligent TCM Inheritance Information Management and Mining Platform for Key Research and Development Projects of the Ministry of Science and Technology(2017YFC1700302).
通信作者:曹存根(cgcao@ict.ac.cn)

进行分类标注时，领域人员有时不易区分粒度细微语义相近的分类标签，使得部分标注结果存在冲突和遗漏。当分类标签数量很多时，这些问题会进一步加剧。同时，拥有多个分类标签（在不引起混淆的情况下，下文将分类标签简称为标签）的文本难以被标全的问题也较为突出。

我们具体分析某省电信公司投诉工单处理流程及涉及的分类标签标注问题。该公司每天接收到数千次投诉电话，线上业务人员根据电话内容编写出一张含有300字左右的投诉工单，然后从1600多种标签中挑选一个或多个标签对其进行标注，并且将标注的工单交给下游业务人员进行后续的业务处理。由于线上业务人员处理业务时间匆忙，同时投诉工单分类标签的数量多、分类标签具有动态性、部分标签不易区分，导致业务人员常常出现错标和漏标。

下面是一个业务人员编写的具体投诉工单。

阜阳受理内容：用户来电称在所在区域信号差，影响通信网络，要求局方尽快处理。核实情况：外省号码无法启用诊断，请后台核实。处理要求：用户要求局方尽快处理。

表1列出了该电信公司业务分类标签体系中的部分标签，这些标签具有多个层次，例如标签“网络质量-移动语音-省际漫游-外省号码省内使用信号差”体现了4个层次，即网络质量问题、移动语音问题、省际漫游问题以及外省号码省内使用信号差问题。

表1 业务分类标签体系中的部分标签

Table 1 Some labels in business classification label system

标签名
网络质量-移动语音-信号弱/不稳定-部分地址信号差
网络质量-移动语音-省际漫游-外省号码省内使用信号差
网络质量-固定电话-国际长途质量-功能正常无法拨打国际长途
网络质量-固定电话-国际长途质量-国际长途提示呼叫限制
渠道服务-网掌厅-客户体验-无法登陆-密码锁定/冻结无法登陆
渠道服务-网掌厅-客户体验-无法登陆-网厅无法登陆
渠道服务-网掌厅-客户体验-无法正常办理业务-证件无法上传
渠道服务-网掌厅-客户体验-无法正常办理业务-页面问题
渠道服务-网掌厅-业务受理-退款不及时-时限内催退款
渠道服务-网掌厅-业务受理-退款不及时-超时退款

该公司业务人员从分类标签体系中选择分类标签“网络质量-移动语音-省际漫游-外省号码省内使用信号差”对上述工单进行标注，这是一个正确的标注。但是，有些业务人员会选择标签“网络质量-移动语音-信号弱/不稳定-部分地址信号差”标注上述工单，这是一个错误的标注。错误的原因是正确的标签（前者）是错误的标签（后者）的具体化，前者更能反映出该工单“外省号码省内使用信号差”的内涵，而后者只反映了“信号差”的内涵，该内容将在第2节进行全面分析。

仅在半年内，该公司新增了3项业务，分别是5G、智慧家庭以及携号转网，包含80多个投诉工单分类标签。包括上述新增的80多个分类标签在内，我们遇到了7次标签新增、细化和修改，共涉及约140个标签，新增业务产生了80多个新标签，细化产生50多个新标签，并相应地修改了2个标签。

多年来，上述问题一直困扰着该电信公司的业务专家和普通业务人员。我们认为，这种困扰的根源就是前面提出的分类标签体系的动态性和不易区分性。

分类标签体系的动态性和标签不易区分性带来的标注冲突和遗漏使得直接使用文本分类方法的效果较差，因此训练文本分类模型前需对标注冲突和遗漏进行检测与修复。我们主要考虑实际业务中的两个需求：1）算法具有一定可解释性；2）算法可以帮助业务人员持续提高标注水平。为此，我们将全自动的标注有噪声时的分类学习方法和全手动的主动学习数据清理方法相结合，提出了一种实用的解决方案，该方案主要包括：

（1）非完美多分类标签体系的概念模型。我们深入分析了实际领域多分类标签体系存在的问题以及人工数据分类标注的不一致性等问题，提出了一种非完美多分类标签体系的概念模型，用于向领域标注人员阐述本文总结的问题以及对应的解决思路。

（2）高质量种子训练集获取方法。根据标注冲突和遗漏的原因，可以将种子训练集分为以下3种：新分类标签体系下标注的数据、针对易混淆类别仔细标注的数据、针对标注人员水平不均衡问题挑选出来的数据。

（3）问题数据检测与修复方法。我们将全自动的标注有噪声时的分类学习方法和全手动的主动学习清理方法相结合，从而对已标注数据中的分类标注冲突和遗漏进行检测，然后进行半自动化修复。因为在现实业务中，具有一定量的标注人员可以基于种子训练集对存在分类标注冲突和遗漏的数据集进行进一步的分析。标注人员虽然需要付出一定的代价，但在一定范围内是可以接受的。其次，领域标注人员比较倾向于采用交互式的方法，以便提高方法的可解释性。

（4）基于BERT[2]的短文本分类学习。针对标签体系具有层次关系的特点，同时学习标签体系和短文本的特征向量。针对种子训练集相对较小的问题，采用固定部分神经网络层次，只训练剩余层次的策略。针对种子训练集之外的新的训练数据集，采用增量训练的策略，从而缩短训练时间。

2 相关工作

2.1 短文本分类

常用的短文本分类方法包括基于专家规则的分类以及基于机器学习的分类方法，如SVM、决策树[3]、KNN和神经网络等。2018年Google提出了基于Transformer[4]的预训练语言模型BERT。该模型使用大量语料进行无监督预训练，BERT提取出的特征具有较好的表示能力和可迁移性，并在自然语言处理的许多任务上取得了当时的最好成绩。

对于分类标签体系中标签较多、标签体现出层次关系的分类任务，研究者提出了许多利用层次关系来改进分类模型的算法。例如，HIAGM[5]（Hierarchy-Aware Global Model）使用TreeLSTM或图神经网络获取标签的向量表示，从而建模标签之间的联系。在获取标签的向量表示之前，HIAGM基于训练集获取了父类标签出现时各子类标签出现的条件概率，以此作为获取标签特征向量的模型的一种输入信息。在获取到标签特征向量后，HIAGM采用一个Attention[4]网络将标签特征向量和使用RNN及CNN获得的文本特征向量融合，以此作为分类的特征向量。

2.2 标注有噪声时的分类学习

分类标注有噪声的情况在使用深度学习方法的过程中

十分常见，标注中的噪声对所训练模型的泛化性影响较大，因此有许多研究者针对标注有噪声时的分类学习方法进行了研究[6]。

在损失函数设计方面，Natarajan 等[7]在损失函数中考虑了预先给定的类别错误率，从而得到了一种能够容忍标注噪声的模型。Bootstrapping 方法[8]通过将神经网络的预测和原标注进行加权平均得到软标注，然后利用软标注进行训练；标注加权平均的系数通过交叉验证获得。

一些方法首先标注少量高质量种子训练集，然后使用元学习方法自动调整样本在损失函数中的权重或者为样本生成软标注[6]。Automatic Reweighting[9]是一种基于训练时模型对样本的梯度学习训练样本权重的元学习方法，它使用一个较小的高质量数据集调整样本在损失函数中的权重，使得训练后的模型在该高质量数据集上的损失函数最小。Zhang 等[10]在利用高质量数据集学习样本权重的同时，也为样本生成软标注。该方法所使用的种子训练集的大小只占整体训练集的 0.2%。Li 等[11]将高质量种子训练集训练的模型作为教师模型，通过知识蒸馏的方法优化标注有噪声的数据训练的学生模型。另外，Li 等还在损失函数中考虑了标签之间的关系，通过引入知识图谱来缓解种子训练集较小带来的问题。

在训练策略方面，MentorNet 方法[12]从一个预训练的教师网络中挑选出损失函数值小的样本作为没有噪声的样本交给学生网络学习。Han 等[13]在 MentorNet 的基础上提出了 Co-teaching 方法，同时训练两个网络，然后互相为对方挑选样本。Chen 等[14]将标注有噪声的数据集进行随机划分，使用交叉验证方式挑选出 loss 较小的样本作为没有噪声的样本；最后，在挑选出的样本上使用 Co-teaching[13]的训练策略进行神经网络训练。

2.3 半自动数据清洗

检测和修复脏数据是数据分析中长期面临的挑战。对于大型数据集，一种有效的策略是人工和机器配合进行半自动的数据清洗。ActiveClean[15]帮助用户进行渐进和迭代地清洗，可清洗的错误类型包括特征及标注中的冲突和遗漏。它将数据清洗建模为小批量随机梯度下降算法 Mini-batch SGD 的变种。具体而言，该算法根据当前的清洗结果和原脏数据的对比结果，采样最具清洗价值的一小批数据，清洗后更新模型，然后继续采样。实验结果表明，在清洗相同数据量时，这种采样算法清洗后的模型的平均精度最多是随机采样清洗的 2.5 倍。

2.4 主动学习

主动学习研究如何挑选更有价值的样本给人工标注，以降低标注的成本。按照数据选择的策略，主动学习主要可分为四大类：基于委员会的方法、基于不确定性的方法、基于多样性的方法和基于期望模型改变的方法[16]。

基于委员会的方法使用多个模型对样本进行投票，选择其中分歧较大的样本进行标注。在构造委员会的方法上，Abe 等[17]使用集成学习方法 Boosting 和 Bagging 来得到多个不同的模型。

基于不确定性的方法定义并测量样本的不确定性，其认为样本不确定性越大则标注价值越高。Yakout 等[18]通过采样部分类别后计算熵来定义不确定性。

基于不确定性的方法仅考虑了单个样本自身的信息，容易产生冗余样本，基于多样性的方法则尽量选择能够代表未标注样本整体分布的样本子集进行标注。Nguyen 等[19]将待标注的数据进行聚类，在每个聚类得到的簇中只采样部分数据进行标注，从而避免重复标注类似的样本。

3 非完美多分类标签体系下的领域短文本分类问题与分析

本节给出了非完美多分类标签体系的概念模型，及其引发的分类标注冲突和遗漏问题的解决思路。

3.1 非完美多分类标签体系的动态性分析

定义 1（标签间的逻辑蕴含关系） 对于两个标签 l_1，l_2，对于任意给定的文本 d，若 d 具有标签 l_1，那么 d 也具有标签 l_2，此时我们称 l_1 逻辑蕴含 l_2，记为 $l_1 \rightarrow l_2$。

例如，投诉工单（在不引起混淆的情况下，下文将投诉工单简称为工单）：

阜阳受理内容：用户来电称在所在区域信号差，影响通信网络，要求局方尽快处理。核实情况：外省号码无法启用诊断，请后台核实。处理要求：用户要求局方尽快处理。

分类标签：网络质量-移动语音-省际漫游-外省号码省内使用信号差。

它对应的标签和“网络质量-移动语音-信号弱/不稳定-部分地址信号差”具有蕴含关系，出现“外省号码省内使用信号差”问题必然也会出现“部分地址信号差”的问题，即“网络质量-移动语音-省际漫游-外省号码省内使用信号差”→“网络质量-移动语音-信号弱/不稳定-部分地址信号差”。

定义 2（多分类标签体系） 分类标签体系 $\Sigma=(L,R)$，其中分类标签集 $L=\{l_1,l_2,\cdots,l_n\}$，$R=\{\Rightarrow,isa,\rightarrow,\infty\}$ 是定义在 L 上的二元关系，其中：

(1)$\Rightarrow$ 表示标签间的混淆关系。若标签 l_1 在分类标注时容易被错分为标签 l_2，则记为 $l_1 \Rightarrow l_2$，我们称 l_1 为错误分出标签，l_2 为错误分入标签。

(2)isa 表示标签内部的上下位关系，上下位关系中没有上位的分类标签也被称为顶层分类标签（简称大类）。例如，在标签“智慧家庭-天翼看家-使用问题-无法使用-终端问题”中，“智慧家庭”和“天翼看家”存在上下位关系即“天翼看家”isa“智慧家庭”，“智慧家庭”是一个大类。

(3)$\rightarrow$ 表示标签之间的蕴含关系，参见定义 1。

(4)∞ 表示标签之间的共现关系。

通常在实际应用中，集合 L 中的标签数量较多，如多于 1 000。同时，在实际使用多分类标签体系的过程中，有些文本可以被标注为多个标签，且多分类标签体系下允许一个子标签有多个父标签。另外，这里定义的多分类标签体系本质上就是一个知识图谱。

定义 3（非完美多分类标签体系） 非完美多分类标签体系是具动态性的多分类标签体系，它的动态性来源于以下 4 点。

(1)不及时：体系中缺乏描述新业务的分类标签集合，包含描述已过时业务的分类标签集合。

(2)不细致：体系中存在需要被拆分或细化的分类标签。

(3)不简要:需要合并一些体系中的分类标签集合,以得到概括性的标签。具体而言,存在一些文本,其需要的标签l'被体系中的多个标签所逻辑蕴含,但l'不在分类标签体系中。

(4)不全面:存在一些文本,其所需要的分类标签因为在当前数据集中较少出现而没有被总结出来,因此不在分类标签体系中。

定义 4(时序分类标签体系) 时序分类标签体系$\Sigma_t=(L,R,T)$是处于某个时间 T 的分类标签体系。

基于这些定义,我们具体分析投诉工单分类任务中的非完美多分类标签体系以及由其引起的数据分类标注问题。

(1)不及时的缺点。在进行工单分析任务的半年中,我们遇到了 3 次业务新增,它们分别是 5G、智慧家庭和携号转网,包含 80 多个标签。记当前分类标签体系为Σ_t,新业务的增加会使得Σ_t下的标注在Σ_{t+1}下出现冲突或遗漏。

(2)不细致而需要拆分或细化的缺点。随着某些标签下的数据积累,标注人员发现了大量的特定问题,然后将这些问题归结到一个专门的子标签。若在Σ_{t+1}下再出现类似内容的工单,则将其标注为专门的子标签。在不增删分类标签体系中的标签时,也可以进行标签的拆分,操作方式是将一个标签对应的部分数据移动到另外一个标签下。标签的拆分或细化会使得Σ_t下的标注在Σ_{t+1}下出现冲突。在工单分析任务中,我们遇到了 7 次标签新增、细化和修改,逐步增加了约 140 个标签(包括上述“关于不及时的缺点”中的 3 次业务新增在内),并相应地修改了 2 个标签。其中,一次标签体系调整发生在业务快速变化时期,导致当时正在服务的调整前训练的模型的精度降低了 5%。

(3)不简要、概括性标签缺失的缺点。标注人员面对概括性标签对应的工单时,往往会从蕴含该概括性标签的分类标签中挑选一个,不同标注人员的选择可能不一致。以“4G 上网信号差”“3G 上网信号差”和“通话信号差”这 3 个标签为例,当工单的内容只包含“手机信号差”,而没有具体表述是“上网”还是“通话”时,标注人员往往在 3 个标签中随意选择一个进行标注,进而使得数据集中出现标注冲突。

(4)不全面的缺点。根据一次标注实验的统计,大约有 5%的工单被标注为内容不明确或没有对应的节点。但是,一些标注人员可能给这些工单标上含义相近但不准确的标签,而非标注为“无标签”。

3.2 分类标签的不易区分性分析

多分类标签体系下的标签的不易区分性主要体现在具有逻辑蕴含关系的标签上,逻辑蕴含关系的左值和右值均可导致标注人员错标。如果标注人员不对文本进行精细的分析,那么就不能给出蕴含关系中语义更加精细的左值标签。当他们不注意判断左值标签中的一些属性能否满足时,就容易将本该标注为右值的工单标注为左值。

例如,第一节展示过的具有不易区分标签的工单样例:

阜阳受理内容:用户来电称在所在区域信号差,影响通信网络,要求局方尽快处理。核实情况:外省号码无法启用诊断,请后台核实。处理要求:用户要求局方尽快处理。

分类标签:网络质量-移动语音-省际漫游-外省号码省内使用信号差。

上述工单对应的标签逻辑蕴含“网络质量-移动语音-信号弱/不稳定-部分地址信号差”,因此导致了“网络质量-移动语音-省际漫游-外省号码省内使用信号差”⇒“网络质量-移动语音-信号弱/不稳定-部分地址信号差”。

也存在蕴含关系中的右值⇒蕴含关系的左值的情况,例如工单:

在饿了么申请了骑士卡 69 元,通话优惠卡,现在激活不了这个卡,营业厅也查不到这个卡的资料,当时是实名申请的卡。

分类标注:4G 业务-4G 套餐业务-IT 类-开通激活-新装号卡激活不成功。

该工单对应的标签被“渠道服务-网掌厅-客户体验-无法正常办理业务-视频认证失败/无法激活”逻辑蕴含,但该工单缺乏关于“网掌厅”的描述,因此不能被分入其中,只能分入更加通用的右值标签“4G 业务”,导致有些标注人员不能正确地对其进行区分。

在拥有多个标签的工单容易漏标的问题上,由于分类标签体系从不同的业务角度刻画问题,同时,一个工单也可能包含多个问题,因此一个工单可能会对应几个标签,但标注人员的记忆有限,因此不易给全所需标注。

例如工单:用户要求取消智能组网业务。用户称营业厅不予取消,要求立即解决。

分类标注:渠道服务-营业厅-解释、说明、宣传不清晰/错误-其他-营业厅业务办理权限”“智慧家庭-全屋 WiFi(智能组网)-业务办理问题-开通/取消-要求开通/取消”。

从服务提供者的角度看,该工单应该被标为“渠道服务-营业厅-解释、说明、宣传不清晰/错误-其他-营业厅业务办理权限”,但从业务产品的角度看,应被标为“智慧家庭-全屋 WiFi(智能组网)-业务办理问题-开通/取消-要求开通/取消”。

从以上对标签不易区分性的 3 方面分析可以看出,面对细粒度的多分类标签体系,标注人员通常会面临很多困难。同时,在不熟悉的标注人员持续理解分类标签体系的过程中,之前因理解不当而错标漏标的工单往往没有办法再次纠正。关于不易区分性对人工分类标注带来的具体困难,我们抽检了包含易混淆标签的数据集,发现其中的原人工标注的精度只有 0.8,召回率更是低至 0.5(见第 5 节)。

3.3 解决思路

我们可以根据导致标注冲突和遗漏发生的原因,在领域业务人员的配合下,得到少量高质量数据集。可以从以下两个重要问题出发:

(1)从分类标签体系的动态性出发,在完善后的新分类标签体系下标注数据。

(2)从分类标签不易区分性出发:

1)针对粒度细微语义相近的标签,单独标注具有混淆关系的标签下的数据。

2)针对标注人员之间对分类标签体系的理解程度不均衡问题,利用分类标注的一些特征,如所属标注人员、标注结果的统计特征等,启发式地挑选标注质量较高的数据。

3)针对多分类标签体系下的漏标问题,采用交叉验证的方法,抽样训练集得到多份子数据集,在此基础上训练多个文本分类模型。若这些模型的集成对已标注样本的预测和原标

注不同,则将它们挑选出来重新标注。

这些高质量的数据集可以作为种子训练集,在种子训练集上的文本分类模型可用来挑出低质量的数据集中可能有问题的样本,这些样本需要被重新标注。

综上所述,解决问题需要进行以下工作:

(1)获取标注质量高的种子训练集,主要可分为 3 种:最新的分类标签体系下标注的数据、针对易混淆标签单独标注的数据,以及根据标注时的特征启发式获取的数据。

(2)利用种子训练集检测与修复低质量的数据集中的标注冲突和遗漏。

4 非完美多分类标签体系下的领域短文本分类分析与学习

本节介绍了非完美多分类标签体系下的领域短文本分类分析与学习方法的总体流程和相关方法,包括对训练数据集[1)]中可能存在的标注冲突和遗漏进行检测与修复的方法、基于 BERT 的两种短文本分类模型。

4.1 总体流程

图 1 给出了非完美多分类标签体系下的领域短文本分类分析与学习的流程。

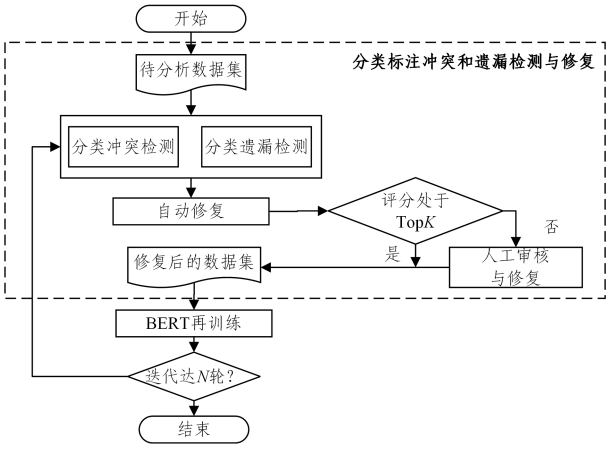

图 1 总体流程图

Fig. 1 Overall flowchart

本文提出的文本分类分析方法是一种半自动化的方法。该方法基于领域标注人员配合构造的分类标注质量高的种子数据集,使用标注有噪声时的分类学习方法训练一个文本分类模型,并将其作为标注冲突和遗漏的检测与修复模型。基于该模型,为待分析数据集中的数据打分,基于打分挑出可能存在标注冲突和遗漏的数据。对于这些数据中得分低的部分,由人工挑选得到修复后的标注;对于得分高的部分,则不经人工挑选,直接将检测与修复模型预测的标注和原标注做一定的融合,从而得到修复后的标注。由上述说明可知,“半自动”体现在两方面:

(1)人工修复时机器帮助挑出值得标注人员检查的标注,以供其选择。

(2)可不经人工挑选,机器直接修复部分标注。

本文提出的文本分类分析与学习方法是一种迭代式的方法。该方法的流程类似于小批量随机梯度下降算法,因此“迭代式”体现在两个方面:

(1)每次处理待分析数据集中的一个子数据集。

(2)对待分析数据集进行多轮分析。

下面具体介绍文本分类分析与学习方法的细节。

4.2 分类标注冲突和遗漏检测与修复算法

4.2.1 设计思路

(1)分类标注冲突和遗漏的检测与修复模型的学习方法选择。从标注有噪声的分类学习的相关工作[11,20]的实验结果中发现,这些改进的方法与直接将种子训练集和标注有噪声的训练集混合后训练的模型的 MAP(Mean Average Precision)最高差 9%,与先在有噪声数据集上预训练然后在高质量种子训练集上微调得到的模型的 MAP 则差 3%左右,而与将预训练和微调得到的这两个模型集成的模型的 MAP 只差 1%左右。综合考虑这些方法的性能和易用性后,我们采取了在有噪声数据集上预训练得到基础模型,在该模型上用种子训练集微调的学习方法。在使用检测与修复模型为文本推断标签时,可与基础模型集成,以进一步提高预测的质量。

(2)全手动的主动学习清理方法选择。从标注有噪声的分类学习的相关工作的实验结果中发现,完全修复的训练集上的模型,其 MAP 最高可比不修复但使用这些标注有噪声的学习方法高 6%。因此,我们还需手工修复一些数据。首先使用基于委员会的方法,将待修复数据集中的原标注当作一个委员模型,将种子训练集上的模型作为另外一个委员模型,选择它们间分歧较大的样本进行标注。此后,根据种子训练集上的模型对样本的不确定度再次进行挑选。

4.2.2 算法流程

为了方便描述,我们先引入一些符号。数据集是由文本及其标注构成的集合,记为 D;训练得到的基于 BERT 的短文本分类模型记为 θ。检测和修复算法的输入是需要修复的数据集 D_t,以及比前者标注质量更高的种子训练集 D_{seed}。检测和修复算法的 3 个阶段如下。

(1)种子训练集获取阶段。在领域标注人员的配合下,获取 D_{seed}。基于 D_t 和 D_{seed},得到标签之间的混淆关系以及容易被错误分入的标签集合 LS_{noise}。

(2)检测与修复模型训练阶段。从 D_t 中采样一个小批量的待修复数据集 S,用 D_t 中剩余的数据 $D_t - S$ 训练基于 BERT 的短文本分类模型 $\theta^{(\text{base})}$,用 D_{seed} 在 $\theta^{(\text{base})}$ 上微调得到检测与修复模型 $\theta^{(\text{detector})}$,简记为 $\theta^{(\text{det})}$。

(3)数据修复阶段。选出 S 中有标注在 LS_{noise} 中的样本构成数据集 S_{noise},使用 $\theta^{(\text{detector})}$ 挑选出 S_{noise} 中可能存在标注冲突和遗漏的数据集 S_{diff}。利用 $\theta^{(\text{detector})}$ 和 LS_{noise} 对 S_{diff} 进行半自动修复,得到 S_{clean} 并加入 D_{seed}。

我们将所提出的检测与修复算法称为 ActiveCleanBERT,其工作流程如图 2 所示。

1) 在本节以及后续章节中,本文将使用数据集、样本集、训练集以及种子训练集等术语。这些数据集通常是由一些短文本及它们的分类标注组成的。当特指本文的具体任务时,这些数据集则指由投诉工单及它们的分类标注组成的数据集合。

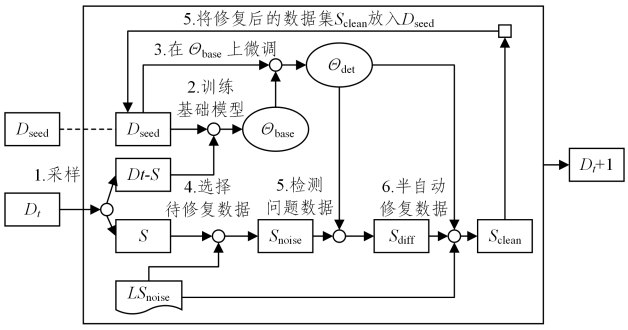

图 2 ActiveCleanBERT 的流程图

Fig. 2 Flowchart of ActiveCleanBERT

接下来将详细介绍算法的实现流程。为使描述简洁，首先将训练、预测然后找原标注和预测结果之间的差异的这一流程定义为函数 findDifference，简写为 F。该函数有 3 个输入，算法步骤如算法 1 和算法 2 所示。

算法 1 $F(D_{base}, D_{finetune}, D_{eval}) \rightarrow D_{eval}^{(diff)}$

输入：D_{base}（用于全量训练的数据集）；$D_{finetune}$（用于全量训练后增量训练的数据集）；D_{eval}（在该数据集上，找原标注和 $D_{finetune}$ 上模型的预测的不同）

输出：可能存在标注冲突和遗漏的数据集

1. 利用 D_{base} 全量训练基于 BERT 的短文本分类模型，得到的模型记为 $\theta^{(base)}$。

2. 利用 $D_{finetune}$ 在 $\theta^{(base)}$ 上进行增量训练，得到的模型记为 $\theta^{(finetune)}$。

3. 利用 $\theta^{(finetune)}$ 对 D_{eval} 进行推理，得到预测的结果数据集 P，该数据集中包含原文本及其预测的分类标注。

4. 对于 D_{eval} 中的样本，若其文本对应的分类标注和该文本在 P 中的分类标注不同，则将这条数据选出来放入数据集 $D_{eval}^{(diff)}$。

5. 返回 $D_{eval}^{(diff)}$。

算法 2 $\text{ActiveCleanBERT}(D_t, D_{seed}) \rightarrow D_{t+1}$

输入：D_t（需要修复的数据集）；D_{seed}：（比 D_t 标注质量更高的种子训练集）

输出：修复后的数据集

1. 若输入的种子训练集 D_{seed} 为空，则通过以下两个步骤来获取：

1.1. 在领域标注人员的配合下，针对分类标签体系的动态性或标签不易区分性获取少量高质量的种子训练集 D_{seed}，包括新分类标签体系下标注的数据和针对易混淆类别专门标注的数据。在针对易混淆标签进行标注时，为使种子训练集能覆盖易混淆标签所属大类下的除易混淆标签本身的其他标签，需根据上下位关系额外标注一定量的数据。

1.2. 如需清理种子训练集本身，根据数据标注时的特征启发式地获取用于修复种子训练集的种子训练集。具体而言，在使用 D_t 上训练的模型协同人工标注 D_{seed} 时，挑选出该模型的预测有错误或遗漏的这部分数据 D_{seed}^{*}。然后，从 $D_{seed} - D_{seed}^{*}$ 中采样和 D_{seed}^{*} 相同数量的数据，和 D_{seed}^{*} 一起作为修复种子训练集的种子训练集 $D_{seed}^{(refiner)}$，$D_{seed} - D_{seed}^{(refiner)}$ 作为被检测和修复的数据集，使用 ActiveCleanBERT 算法进行修复，即：

$\text{ActiveCleanBERT}(D_{seed} - D_{seed}^{(refiner)}, D_{seed}^{(refiner)}) + D_{seed}^{(refiner)} \rightarrow D_{seed}^{t+1}$。

2. 基于 D_t 和 D_{seed} 获取需要修复的标签集合 LS_{noise}。记当前数据集为 D_t，将 D_{seed} 划分为 $D_{seed}^{(train)}$ 和 $D_{seed}^{(test)}$ 两个数据集，从 $F(D_t, D_{seed}^{(train)} + D_t, D_{seed}^{(test)})$ 中统计出混淆关系，选出错误分入频次较高的标签构成 LS_{noise}。

3. 小批量迭代式地检测和修复数据集。当 $D_t \neq \emptyset$ 时，执行：

3.1. 从 D_t 中采样一个小批次的待修复数据集 S，从中选取有标注在 LS_{noise} 中的样本，构成数据集 S_{noise}。这个选取的操作记为 T，即 $S_{noise} = T(S, LS_{noise})$。

3.2 错误探测与修复模型的训练和预测：

使用 $D_t - S$ 及 D_{seed} 训练 BERT 得到 $\theta^{(detector)}$。对于 S_{noise} 中的文本，若 $\theta^{(detector)}$ 对它的预测标注和它在 S_{noise} 中的原标注存在不同，则将它挑出来，即执行 $S_{diff} = F(D_t - S, D_{seed}, S_{noise})$。

使用打分函数 V 对 S_{diff} 中的文本进行打分。我们使用熵来定义打分函数 V(d)，对于一个文本 d，$\theta^{(detector)}$ 给出了其属于各个标签的后验概率，筛选出其中概率大于阈值的 n 个标注，其概率分别为 $p_1, \cdots, p_n, \varepsilon = 1 \times 10^{-8}$，则：

$$V(d) = \frac{\sum_{i=1}^{n} p_i \log p_i}{\log \frac{1}{n} + \varepsilon}$$

3.3. 半自动地修复数据。选取所得分数不在 top-k 中的文本，将其在 S_{diff} 中的原标注和 $\theta^{(detector)}$ 预测出的标注融合，得到提供给人工审核的建议标注。将建议标注和文本组成数据集 S'_{diff} 交由人工审核得到 S'_{clean}；对于 S_{diff} 中分数在 top-k 中的剩余文本，可直接用融合的标注替代其在 S_{diff} 中的原标注。最后，将半自动修复后的 S_{diff} 放回 S 中得到 S_{clean}，即对于 S 中的文本，若其在 S_{diff} 中，则用半自动修复后的标注替换其在 S 中的原标注。标注融合和人工审核方法的具体细节如下：

融合文本的原标注和 $\theta^{(detector)}$ 预测出的标注得到建议标注的具体方式为，删除原标注中在 LS_{noise} 中的标注，对剩余的和 $\theta^{(detector)}$ 预测出的标注取并集。标注融合的必要性在于，种子训练集只在具有混淆关系的标签上比待修复数据集质量高。

在人工审核时，为了突出需要关注的标签，我们在操作界面上同时给出了原标注和建议标注，请标注人员挑出其中的错标和漏标。另外，为了减少标注人员审核时需要补充的标注，若一个文本的建议标注为空，则将 $\theta^{(detector)}$ 预测出的后验概率最高的标签作为建议标注。

3.4. 数据集更新，即 $D_t = D_t - S$，$D_{seed} = D_{seed} + S_{clean}$。

4. 将 D_{seed} 并入 D_t 得到 D_{t+1}。并入操作指对于 D_{seed} 和 D_t 中的文本集合的交集，用 D_{seed} 中的标注替代 D_t 中的标注放入结果数据集 D_{t+1} 中；对于在 D_{seed} 而不在 D_t 中的文本，将它和它在 D_{seed} 中的标注放入结果数据集 D_{t+1}。至此，算法完成了对待修复数据集的一轮迭代，返回 D_{t+1}。

综上所述，在算法技术上，我们针对分类标签体系的动态性和标签不易区分性提出了 3 种种子训练集获取方法，其中第三种获取方法比较特别。在利用种子训练集的方法上，我们先用标注有噪声时的分类方法训练检测和修复数据的模型，基于此使用主动学习挑出困难样本，请业务人员帮助进行手工标注。具体而言，在标注有噪声的分类学习方面，本文方法类似于 INCV[20]，Bootstrapping[8]，Data coefficients[10]，以及 Li 等[11]提出的方法。INCV 将标注有噪声的数据集随机划分，使用交叉验证或异常检测的方式挑选出损失大的样本，从而将没有噪声的样本挑出来，然后在这些样本上进行训练。通过随机划分数据集，利用交叉验证推断出不确定性高的样本，但对于部分样本，我们会对其进行人工修复而非丢弃它。Bootstrapping 和 Data coefficients 方法为带噪声的训练样本生成软标注，本文利用 $\theta^{(detector)}$ 和标签间的混淆关系做自动

修复的方法与其类似。Li 等利用知识图谱弥补了高质量种子训练集较小的缺陷，我们也利用标签之间的关系来缩小检测和修复的分类标注的范围。在主动学习方面，本文应用了基于委员会[17]的和基于不确定性的[18]数据选择方法。

4.3 领域短文本分类学习方法

本文基于常用的文本特征提取模型 BERT-base 搭建文本分类模型。在训练 BERTTextCls 以及使用它进行推理之前，需对输入文本做预处理。文本预处理主要是去除文本中的空白符，包括换行符、制表符以及空格，然后截取文本中一定量的字符。训练时，将一条带多个标注的训练数据展开成多条单标注数据，然后使用交叉熵损失函数进行训练。预测时，模型给出领域文本属于各标签的概率 $p(L|d)$，然后保留概率大于选定阈值的预测标签。

除了基础的 BERTTextCls 模型，我们还参考了多标签分类算法 HIAGM[5]，使用 TreeLSTM 提取具有层次关系的分类标签体系的特征，与待分类文本的特征一起进行文本分类任务的训练。关于待分类文本的特征提取，以往的 HIAGM 原文中使用的 TextRCNN 方法已经比较陈旧，这里使用较新的 BERT 方法，我们将这种文本分类模型记为 BERTTextCls＋HIAGM。

5 实验与分析

本节首先通过实验分析 BERTTextCls 在投诉工单领域短文本分类任务上的性能，然后验证标注检测和修复算法 ActiveCleanBERT 的有效性，实验分为两个部分进行：处理非完美多分类标签体系调整的实验（简称实验 1）、处理人工标注困难的实验（简称实验 2）。

5.1 领域短文本分类模型的实验

为分析分类模型对领域短文本分类任务的效果的影响，本文选取了常用的文本分类模型，这些分类模型的精度如表 2 所列。实验的数据集中包含 40 000 条数据，将其中的 4 000 条作为测试集，其余作为训练集。

表 2 不同文本分类方法的精度

Table 2 Accuracy of different text classification methods

文本分类方法	精度
使用词特征的 SVM	0.670
使用字和字的 bigram 特征的 SVM[21]	0.690
没有预训练的 BERTTextCls	0.700
BERTTextCls	0.799
BERTTextCls＋HIAGM[5]	0.782

从表 2 可以看出，在投诉工单领域短文本分类任务上，BERT 预训练对精度的提升可达 10%左右，而模型结构对精度的影响只在 2%以内。另外，在 BERT 上使用多标签分类算法 HIAGM 后，精度降低了 0.017。这可能是因为 HIAGM 中包含了图神经网络 TreeLSTM，它和 BERT 中的 Transformer 差别较大，从而需要针对性地对其进行细致的参数调整，而不能沿用训练 BERT 时的参数。

为分析模型性能随标注数据积累的提升趋势，我们对比了数据积累的 3 个阶段中 BERTTextCls 模型的性能，3 个阶段的数据量分别为 40000，100000 和 200000。在召回 90% 的投诉工单的情况下，这 3 个阶段下 BERTTextCls 的 F1 值分别为 0.76，0.84 和 0.90。

5.2 实验 1：处理非完美多分类标签体系调整的实验

本实验使用 ActiveCleanBERT 算法处理“天翼看家”类别细化导致的标注冲突和遗漏，该类别原包含 19 个标签，现在向其中添加 8 个标签。实验中标注 668 条种子数据集，按照本文 4.3.2 节中介绍的算法流程对 D_l 分 3 个小批次进行一次迭代。

为更直观地理解标签间的混淆关系，表 3 列出了此次调整带来的部分混淆标签对。

表 3 实验 1 中的部分混淆标签对

Table 3 Some confusing label pairs in experiment 1

正确标签	错误预测
智慧家庭-天翼看家-装维服务-开通类-未装报竣	智慧家庭-天翼看家-其他-要求安装/取消
智慧家庭-天翼看家-使用问题-无法使用-终端问题	智慧家庭-天翼看家-使用问题-无法使用-监控无法使用
智慧家庭-天翼看家-装维服务-开通类-长时间未装	智慧家庭-天翼看家-其他-要求安装/取消

(1)评测方法

由于分类标签体系中的标签具有不易区分性，在来自人机协同标注的数据集上进行评测时，其结果会偏向人机协同标注时使用的模型。因此，我们通过人工对比审核的方式来进行模型性能的比较。

(2)评测结果

表 4 列出了此次实验的总体测评结果。从表中可以看到，数据修复将模型在“天翼看家”类别上的 F1 值提高了 0.037。

表 4 实验 1 迭代流程的总体评测

Table 4 Summarization of evaluations in experiment 1

训练集大小	测试集大小	S_{noise} 总量	人工审核量	原人工标注 F1 值	F1 值提升
200000	124	620	261	0.610	＋0.037

5.3 实验 2：处理人工标注困难的实验

本实验验证了 ActiveCleanBERT 算法处理人工标注困难的有效性。实验使用的种子训练集包括：针对易混淆类别特别标注的数据、根据数据标注时的特征启发式地获取的数据。

5.3.1 ActiveCleanBERT 算法的执行流程

(1)获取种子训练集 D_{seed}

在 θ^{D_l} 的帮助下，针对错误率高的大类，标注一批数据 D_{seed}，包括大类 4G 业务、流量、智慧家庭和渠道服务，共 6500 条左右。

(2)获取混淆关系

基于 D_{seed} 统计混淆关系，取混淆关系中错分入最多的 20 个标签作为 LS_{noise}，LS_{noise} 涉及的大类相比 D_{seed} 收集时用的大类多一个“网络质量”。

(3)修复数据集

按照本文 4.3.3 节中介绍的方法，先用一个小批次修复

种子训练集D_{seed}本身可能存在问题的部分数据集$D_{seed}^{(noise)}$，然后用一个小批次修复D_t中的1/3，记为$D_t^{(part1)}$，最后用一个小批次修复$D_t - D_t^{(part1)}$。

5.3.2 ActiveCleanBERT算法的执行效果评测和分析

(1)迭代流程的总体评测

为从总体上确认实验2的效果，我们总结了修复第三个小批次过程中的关键指标，如表5所列，只评测第三个小批次是因为前两个小批次中修复的审核人员不如第三个小批次的人员熟练。

表5 实验2迭代流程的总体评测

Table 5 Summarization of evaluations in experiment 2

训练集大小	测试集大小	S_{noise}总量	人工审核量	原人工标注F1值	F1值提升
200000	6600	65000	2565	0.650	+0.012

(2)迭代流程中的细节评测与分析

为了确定ActiveCleanBERT算法能检测出数据中的冲突和遗漏，我们基于迭代过程中人工审核过的数据集进行评测，所得的F1值如表6所列。从该表中可以看出，原标注在再次审核后F1值不到0.70。

表6 实验2数据修复过程中各类标注的F1值

Table 6 F1-score of class labels during data repairing in experiment 2

待修复数据集	审核量	标注类型	F1值
$D_{seed}^{(noise)}$	85	原标注	**0.43**
		重打标	0.78
		原标注+重打标	0.80
$D_t^{(part1)}$	1593	原标注	**0.63**
		重打标	0.79
		原标注+重打标	0.87
$D_t - D_t^{(part1)}$	2565	原标注	**0.65**
		重打标	0.61
		原标注+重打标	0.79

为了确定根据数据标注时的特征启发式地获取的种子训练集有效，继续基于表6对修复$D_{seed}^{(noise)}$的过程进行分析。重打标的F1值达到0.78，比原标注高0.35，说明了这类种子训练集的有效性。但由于本次实验是我们初次尝试用该方法检测和修复种子数据集本身，我们期望检测具有较高的精度，因此只人工审核了$D_{seed}^{(noise)}$中的7%的数据，导致$D_{seed}^{(noise)}$中可能还有一些标注冲突或遗漏没有被检测出来。

为了确定数据修复在解决的问题是人工标注困难带来的问题，我们统计了修复过程中主要处理的几个标签在修复前后的指标，结果如表7所列。具体而言，表7列出了在修复$D_t - D_t^{(part1)}$过程中，修复算法主要处理了的因蕴含关系导致混淆的两个标签对中的4个标签的频次，以及在人工审核过的数据集S'_{clean}上，各类标注在这些标签上的指标。这两个具有蕴含关系的标签对为："智慧家庭-费用问题-计费差错-未告知收费不认可"→"智慧家庭-费用问题-用户否认使用-否认开通/不认可费用"和"4G业务-4G数据卡业务-网络质量类-无信号"→"4G业务-4G数据卡业务-网络质量类-网速慢-基站或信号等非终端问题"。观察表7中原标注的指标情况，可发现修复前蕴含关系中的左值标签"智慧家庭-费用问题-计费差错-未告知收费不认可"和"4G业务-4G数据卡业务-网络质量类-无信号"的F1值不到0.7。

表7 各类标注在修复过程中主要处理的几个标签上的F1值

Table 7 F1-score of class labels for hot classes during data repairing in experiment two

标签名	原标注	重打标	标签频次
智慧家庭-费用问题-计费差错-未告知收费不认可	**0**	0.74	30
智慧家庭-费用问题-用户否认使用-否认开通/不认可费用	0.82	0.49	38
4G业务-4G数据卡业务-网络质量类-无信号	**0.67**	0.51	65
4G业务-4G数据卡业务-网络质量类-网速慢-基站或信号等非终端问题	0.61	0.67	45

(3)数据修复对模型的影响的评测和分析

为了确定数据修复能给模型带来提升，并了解人工修复和机器修复分别带来的提升，我们对比了修复前后的性能，对比结果如表8所列。

表8 实验2中ActiveCleanBERT对模型性能的提升

Table 8 Improvement of model with ActiveCleanBERT in experiment 2

修复后的数据集	测试集上的整体提升
$D_{recheck}$	+0.0034精度，+0.0070召回，+0.0051 F1值
D_{RWR}	+0.0080精度，+0.0140召回，+0.0117 F1值

对比实验的细节如下：首先在线上6天产生的约20000条数据中，选取有标签在五大类下的约6500条数据作为测试集，在该测试集上对比第三个小批次修复前后模型的性能。实验的基线为$D_t + D_{seed}$清理了$D_{seed}^{(noise)}$和$D_t^{(part1)}$得到的数据集(约200000条)，为便于表述，将其记为D_{raw}。在D_{raw}之上并入2565条人工修复的数据集，记为$D_{recheck}$；在$D_{recheck}$之上并入2565条机器修复的数据集，记为D_{RWR}(RWR是Recheck-WithReplacing的缩写)。

从表8中可以看到，人工修复和机器修复分别将模型在五大类标签上的F1值提高了0.0051和0.0066，总共提高了0.0117。由于包含五大类标签的数据集的大小约占整体的1/3，因此整体任务的F1值提高了0.004。

为了确定数据修复是通过处理人工标注困难问题而给模型带来了提升，我们统计了表7中两个因蕴含关系而产生混淆的标签在数据修复前后的指标情况，结果如表9所列。从表9可以看出，修复后的模型在蕴含关系中的左值标签上的F1值均提升了0.2以上。这与表7中展示的数据修复将原标注在左值标签上的F1值提升了0.4以上对应。

表9 实验2中部分标签在修复前后的F1值

Table 9 F1-socre of some labels before and after repairing in experiment 2

标签名	D_{raw}	D_{RWR}	标签频次
智慧家庭-费用问题-计费差错-未告知收费不认可	**0.37**	**0.83**	13
智慧家庭-费用问题-用户否认使用-否认开通/不认可费用	0.88	0.83	14
4G业务-4G数据卡业务-网络质量类-无信号	**0.60**	**0.85**	13
4G业务-4G数据卡业务-网络质量类-网速慢-基站或信号等问题	0	0.28	4

结束语 本文结合实际投诉工单业务，系统地分析了多分类标签体系的动态性以及其中标签的不易区分性，从而提出了一种非完美多分类标签体系的概念模型。基于该概念模型，本文提出了一种在领域标注人员配合下构造高质量种子训练集，然后用其检测与修复低质量数据集中的标注冲突与遗漏的半自动方法。该方法结合了全自动的标注有噪声的分类学习方法，以及全手动的主动学习数据清理方法。实验数据证明了所提方法的有效性。

在之后的工作中，可从3个方面改进检测与修复算法。

(1)进一步分析与标注冲突有关的特征，如标注人员的历史标注记录和其标注质量间的联系。

(2)在标注有噪声的分类学习方法上，尝试虽然复杂但效果更好的方法。

(3)在主动学习的方法上，使用基于样本多样性的方法来进一步减少重复标注。

最后，本文提出的概念模型及文本分类方法不依赖特定的领域，后续可进行其他领域的实验研究。

参 考 文 献

[1] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning--based text classification: a comprehensive review[J]. ACM Computing Surveys(CSUR), 2021, 54(3): 1-40.

[2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[3] ZHU Y, TING K M, ZHOU Z H. Multi-label learning with emerging new labels[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(10): 1901-1914.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. MIT Press, 2017: 5998-6008.

[5] ZHOU J, MA C, LONG D, et al. Hierarchy-aware global model for hierarchical text classification[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 1106-1117.

[6] SONG H, KIM M, PARK D, et al. Learning from noisy labels with deep neural networks: a survey[J]. arXiv: 2007.08199, 2020.

[7] NATARAJAN N, DHILLON I S, RAVIKUMAR P, et al. Learning with noisy labels[C]// Advances in Neural Information Processing Systems. MIT Press, 2013: 1196-1204.

[8] REED S, LEE H, ANGUELOV D, et al. Training deep neural networks on noisy labels with bootstrapping[J]. arXiv: 1412.6596, 2014.

[9] REN M, ZENG W, YANG B, et al. Learning to reweight examples for robust deep learning[C]// Proceedings of the International Conference on Machine Learning. JMLR, 2018: 4334-4343.

[10] ZHANG Z, ZHANG H, ARIK S O, et al. Distilling effective supervision from severe label noise[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 9294-9303.

[11] LI Y, YANG J, SONG Y, et al. Learning from noisy labels with distillation[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 1928-1936.

[12] JIANG L, ZHOU Z, LEUNG T, et al. Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels[C]// Proceedings of the International Conference on Machine Learning. JMLR, 2018: 2304-2313.

[13] HAN B, YAO Q, YU X, et al. Co-teaching: robust training of deep neural networks with extremely noisy labels[J]. arXiv: 1804.06872, 2018.

[14] WANG Y, LIU W, MA X, et al. Iterative learning with open-set noisy labels[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 8688-8696.

[15] KRISHNAN S, WANG J, WU E, et al. Activeclean: interactive data cleaning for statistical modeling[J]. Proceedings of the VLDB Endowment, 2016, 9(12): 948-959.

[16] SETTLES B. Active learning literature survey [J]. Science, 1995, 10(3): 237-304.

[17] ABE N. Query learning strategies using boosting and bagging [C]// Proceedings of the Fifteenth International Conference on Machine Learning. JMLR, 1998: 1-9.

[18] YAKOUT M, BERTI-ÉQUILLE L, ELMAGARMID A K. Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes[C]// Proceedings of the 2013 International Conference on Management of Data. ACM Press, 2013: 553-564.

[19] NGUYEN H T, SMEULDERS A W M. Active learning using pre-clustering[C]// Proceedings of the International Conference on Machine Learning. JMLR, 2004: 623-630.

[20] CHEN P, LIAO B B, CHEN G, et al. Understanding and utilizing deep neural networks trained with noisy labels[C]// Proceedings of the International Conference on Machine Learning. JMLR, 2019: 1062-1070.

[21] LI J, SUN M, ZHANG X. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. ACL, 2006: 545-552.

LIANG Haowei, born in 1996, postgraduate. His main research interests include natural language processing and deep learning.

CAO Cungen, born in 1964, Ph. D, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include large-scale knowledge processing and machine learning.

(责任编辑：何杨)