基于弱标签争议的半自动分类数据标注方法

李自强[1],杨 薇[2*],杨先凤[2],罗 林[3]
(1. 四川师范大学影视与传媒学院,四川成都 610066;2. 西南石油大学计算机与软件学院,四川成都 610500;
3. 泰豪软件股份有限公司成都研发中心,四川成都 610041)

摘 要: 当前,深度主动学习(Deep Active Learning,DAL)在分类数据标注工作中获得成功,但如何筛选出最能提升模型性能的样本仍是难题. 本文提出基于弱标签争议的半自动分类数据标注方法(Dispute about Weak Label based Deep Active Learning,DWLDAL),迭代地筛选出模型难以区分的样本,交给人工进行准确标注. 该方法包含伪标签生成器和弱标签生成器,伪标签生成器是在准确标注的数据集上训练而成,用于生成无标签数据的伪标签;弱标签生成器则是在带伪标签的随机子集上训练而成. 弱标签生成器委员会决定哪些无标签数据最有争议,则交给人工标注. 本文针对文本分类问题,在公开数据集 IMDB(Internet Movie DataBase)、20NEWS(20NEWSgroup)和 chnsenticorp(chnsenticorp_htl_all)上进行实验验证. 从数据标注和分类任务的准确性 2 个角度,对 3 种不同投票决策方式进行评估. DWLDAL 方法中数据标注的 F_1 分数比现有方法 Snuba 分别提高 30.22%、14.07% 和 2.57%,DWLDAL 方法中分类任务的 F_1 分数比 Snuba 分别提高 1.01%、22.72% 和 4.83%.

关键词: 深度主动学习;文本分类;伪标签生成器;弱标签生成器;投票委员会
基金项目: 国家自然科学基金(No.61802321);四川省科技厅重点研发计划(No.2020YFN0019)
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112(2024)08-2891-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230648

The Semi-Automatic Classification Data Labeling Method Based on Dispute About Weak Label

LI Zi-qiang[1], YANG Wei[2*], YANG Xian-feng[2], LUO Lin[3]
(1. *College of Movie and Media*, *Sichuan Normal University*, *Chengdu*, *Sichuan* 610066, *China*;
2. *School of Computer Science and Software Engineering*, *Southwest Petroleum University*, *Chengdu*, *Sichuan* 610500, *China*;
3. *Chengdu R&D Center*, *Tellhow Software*, *Chengdu*, *Sichuan* 610041, *China*)

Abstract: At present, deep active learning (DAL) in the classification data labeling work has achieved outstanding success. How to select samples to improve the performance of models is still a difficult problem in deep active learning. We proposes a semi-automatic classification data labeling method based on weak label dispute (Dispute about Weak Label-based Deep Active Learning, DWLDAL). The method iteratively selects samples that is difficult for model to distinguish, and manually annotate these sample. This method contains pseudo label generator and weak label generator, pseudo label generator is trained on accurately annotated datasets to generate pseudo label for unlabeled data; weak label generator is trained on random data subset with pseudo labels. Weak label generator committee are used to determine which unlabeled data is the most controversial and should be manually annotated. We conducted experimental validation on the common datasets IMDB (Internet Movie Database), 20NEWS (20NEWSgroup), and chnsenticorp (chnsenticorp_htl_all) to address the issue of text classification. Three different voting decision-making methods are evaluated from the perspective of the accuracy of data annotation and classification tasks. The F_1 score of data annotation in DWLDAL method is 30.22%, 14.07% and 2.57% higher than that in the existing method Snuba, respectively. The F_1 score of classification task in DWLDAL method is 1.01%, 22.72% and 4.83% higher than that in Snuba method, respectively.

Key words: deep active learning; text classification; pseudo label generator; weak label generator; balloting committee

收稿日期:2023-07-11;修回日期:2024-01-02;责任编辑:梅志强
*通讯作者:杨薇

Foundation Item(s): National Natural Science Foundation of China (No.6180232); Key Research and Development Program of Science and Technology Department of Sichuan Province (No.2020YFN0019)

1 引言

随着深度学习在自然语言处理、计算机视觉、推荐系统等多个领域应用的普及,深度学习模型[1,2]越来越依赖大量准确标记的训练数据集.很多应用场景中的海量数据获取容易,但人工标注成本高.因此,自动和半自动数据标注成为各领域深度学习应用任务的瓶颈.并非每个样本对模型是同等重要,因此,哪些样本最应该标注是核心问题.

数据标注相关方法大致分为半监督学习方法、主动学习(Active Learning,AL)和深度主动学习(Deep Active Learning,DAL)[3].半监督学习方法利用少量有标签数据训练1个模型,利用模型对大量无标签数据进行标注.Trust等人[4]为识别有关经济政策不确定性(Economic Policy Uncertainty,EPU)的新闻文章,提出基于弱监督和深度学习的数据标注方法,在一定程度上取代了手动标记.Ratner等人[5]提出围绕数据编程范式(Data Programming paradigm,DP)而构建的系统Snorkel,用于快速创建、建模并管理用于机器学习的训练数据集.实验证明,该系统在平均预测性能上比之前的启发式方法表现更好.Varma等人[6]提出了用少量标记数据自动生成启发式标注规则的方法Snuba,在弱监督条件下将标签分配给未标记数据.该方法采用多次迭代、单次分步骤的思想实现整体流程,主要分为启发式标注规则生成(Heuristic Generation,HG)、修剪器(Pruner)、验证(Verifier)3部分,其性能比用户定义启发式算法更好.

AL是从无标签数据中选择最有价值的样本进行人工标注的方法,其目的是对尽可能少的、高质量的样本进行标注进而提升模型性能.Park等人[7]提出基于分布感知的主动学习方法,该方法捕获并缓解标记数据集和未标记数据集之间的分布差异,应对过拟合现象.实验结果表明,该方法相对现有的主动学习方法,整体性能有所提升.Chen等人[8]针对电子医疗记录数据需要大量标注样本的问题,提出一种不确定抽样AL方法与基于支持向量机的医疗文本表型算法结合的标注方法.该方法提高了表型分类器性能,且需要更少的标注样本.Goudjil等人[9]提出主动学习文本分类方法,在不影响分类准确性的情况下,利用支持向量机分类器提供的后验概率选择样本供专家标注.实验表明,该方法显著减少标注工作量的同时,提高了分类精度.

DAL近年来在数据标注领域受到越来越多的关注,在保持性能水平的同时大幅度降低标注成本.具体实现过程是利用少量标记数据S训练获得初始深度学习模型,未标记数据集T的样本通过深度学习模型提取特征表示.根据查询策略选择样本进行人工标注,形成新的已标记数据集S',利用S'继续训练深度学习模型,同时更新T.不断重复上述过程,直到达到预定的终止条件.Buchert等人[10]为了在初始阶段选择更大信息量的数据,为主动学习方法引入基于多样性的采样方法,同时提出新的主动学习查询策略.实验结果表明,该方法在2个公共数据集上的性能优于其他主动学习方法.Zhou等人[11]提出半监督学习算法也称为主动深度网络(Active Deep Network,ADN),利用无向图模型构造情感分类器.在半监督学习框架中采用不确定性的查询策略筛选出应标记的样本,再用选定的标记数据和所有未标记数据共同训练ADN体系结构.Talukder等人[12]提出用于新闻准确性检测任务(即识别新闻中的误导和虚假信息)的人机协作学习系统.该系统利用有限的数量标注样本比完全监督学习少1~2个数量级.同时,他们提出了混合查询策略选出最有价值的样本,设计的深度学习模型采用较少数量的局部滤波器,从较小的相关批次样本中进行有效学习.

研究表明,查询策略的设计对数据标注方法整体性能[13]至关重要.一些传统查询策略难以选出最有价值的样本,导致降低人工标注成本难.在提升数据标注的质量方面,很多深度学习模型利用RNN(Recurrent Neural Network)、CNN(Convolutional Neural Network)、LSTM(Long Short-Term Memory)和GRU(Gate Recurrent Unit)等神经网络[14,15]捕获数据中丰富的语义特征.因此,如何设计深度学习模型,让其与查询策略更好的融合是重要问题.本文提出基于弱标签争议的半自动分类数据标注方法,主要贡献如下:

(1)利用深度模型在准确标注的数据集上训练伪标签生成器(Pseudo Label Generator,PLG),给无标签数据打上伪标签.

(2)随机选择多个带伪标签的无标签数据子集,训练弱标签生成器(Weak Label Generator,WLG).根据多个弱标签生成器在无标签数据上的决策争议筛选出最值得人工标注的样本.

(3)构造了弱标签生成器委员会的联合损失函数,保证伪标签生成器和弱标签生成器能体现数据的总体分布,使模型迭代过程收敛.

2 方法介绍

本文提出的基于弱标签争议的半自动分类数据标注方法,是1个迭代式的交互训练过程,主要由5个

核心部分组成,包括:少量有标签数据(Labeled Data,LD)S、大量无标签数据(Unlabeled Data,UD)T、伪标签生成器、弱标签生成器委员会、查询策略. 基于弱标签争议的半自动分类数据标注方法的整体流程如图1所示.

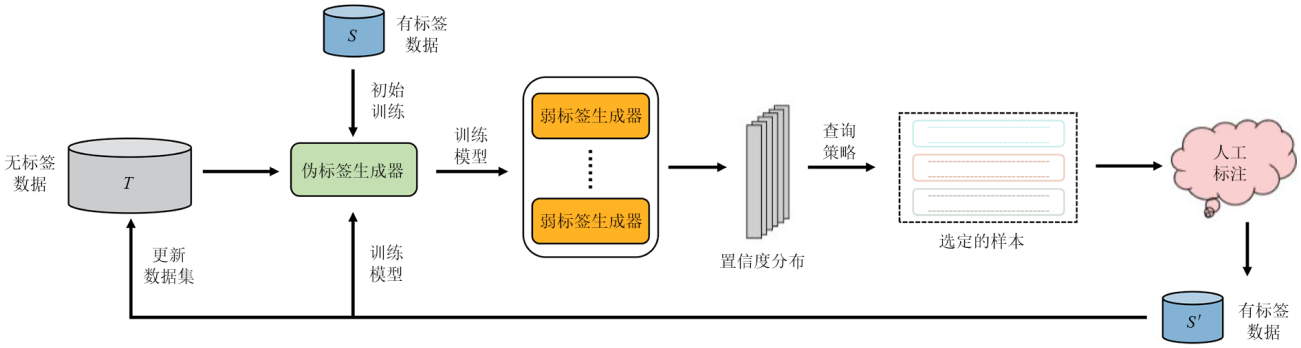

图1　基于弱标签争议的半自动分类数据标注方法

(1)通过少量有标签数据训练伪标签生成器,生成大量无标签数据的伪标签.

(2)随机选择多个带伪标签的无标签数据子集,训练多个弱标签生成器,根据预测结果计算得出多种置信度分布.

(3)根据样本筛选策略,从候选样本中筛选出最有价值的无标签数据样本.

(4)将筛选出的无标签样本进行人工标注,且将标注后的数据用于模型更新训练.

(5)不断迭代进行上面4个步骤,直到满足用户设定的终止条件或无标签数据集合为空.

2.1　伪标签生成器

在标注过程中,少量有标签数据S拥有可信度最高的标签. 利用S训练深度神经网络模型,可以得到在S上可靠度较高的分类模型. 由于深度学习网络模型对准确标注的数据集有依赖性,且无标签数据T的样本数量较大,使该分类模型应用在数据集T上的输出标签可信度不高,称为伪标签生成器.

针对文本分类问题,伪标签生成器整体分为3个阶段,分别是生成文本向量、特征提取和分类,模型框架如图2所示. 首先,使用基于Transformer的双向编码器[16](Bidirectional Encoder Representations from Transformers,BERT)生成文本向量,设样本中的单个句子由n个词组成,$(\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_n)$表示输入的文本字向量. 单个文本字向量由本身字向量(Token Embeddings,TE)、句子的分段向量(Segment Embeddings,SE)、位置向量(Position Embeddings,PE)3部分组成. 传统文本字向量仅考虑了字本身和位置信息,缺少文本的上下文信息. 双向Transformer编码器同时考虑当前单词的上下文和位置信息,将文本字向量经过双向Transformer的编码模块(Trm)处理得到文本的最终向量表示$(\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_n)$.

将BERT预训练模型得到的文本向量作为输入,依次经过以下操作,BatchNorm表示批量归一化网络层,Full connected layer表示全连接网络层. Multiply表示矩阵点乘层,可引入更多的非线性关系,提高模型的表达能力和分类性能. GRU在捕捉长序列语义关联时,能有效抑制梯度消失或爆炸,效果优于传统RNN. 而Bi-GRU[17](Bidirectional Gated Recurrent Unit)在GRU的基础上添加了反向输入,即拥有2个GRU,1个从前往后扫描整个序列,另1个从后往前扫描序列. 基于此结构,BiGRU使模型能同时考虑过去和未来的信息,更加准确地捕获序列数据中的特征信息. 末端的分类器采用Softmax网络层.

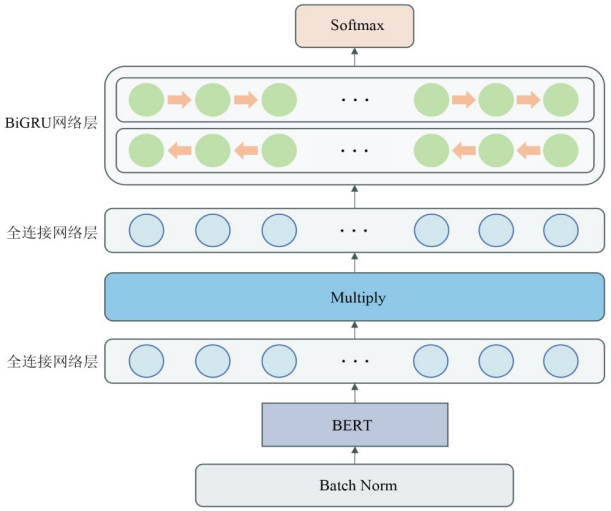

图2　伪标签生成器模型

2.2　弱标签生成器

为了增强迭代过程中的收敛稳定性,本文提出构建弱标签生成器(Weak Label Generator,WLG). 伪标签生成器将大量无标签数据T作为输入,产生T的伪标签. 弱标签生成器与伪标签生成器结构基本相似,并使用T及其伪标签进行训练. 针对文本分类问题,在其末端增加了2个全连接层和1个dropout网络层,整体架构

也分为生成文本向量、特征提取和分类3部分. 将经过BERT模型预训练处理的文本表示作为输入,依次经过批量归一化网络层、全连接网络层、矩阵点乘操作和BiGRU获得文本的特征信息,再经过2个全连接层得到更加丰富的特征信息. 深度神经网络从数据集中学习到不同特征的同时也会学习到噪声,导致出现过拟合现象. 于是,在2个全连接层中引入1个dropout网络层,Softmax网络层将特征信息作为输入,得到分类输出.

2.3 弱标签生成器委员会

基于多个模型委员会预测结果可以提高整体模型的准确性并减少单个模型的偏差,本文设计一种弱标签生成器委员会[18],具体流程如算法1所示. 为避免对大规模无标签数据进行全量训练,采用随机采样思路,选择少量子集进行训练,尽可能拟合数据总体. 但在训练过程中,为了同时提升伪标签生成器和弱标签生成器的性能,将弱标签生成器委员会的联合损失函数定义为多个分类器损失函数的线性组合. 因此,每个成员分类器实际上由伪标签生成器和弱标签生成器组成. 少量有标签数据$S=\{(X_S^i,Y_S^i)\}(i=1,2,3,\cdots,L)$,$X_S{}^i$是第$i$个有标签样本数据,$Y_S^i$是相应的准确标签. 大量无标签数据$T=\{X_T^i\}(i=1,2,3,\cdots,H)$,$X_T^i$是第$i$个无标签样本数据. S和T在数据量方面差距很大,考虑到数据量对模型训练过程产生的影响,从T中随机抽样M个数据子集. T子集集合$A=\{T^j\}(j=1,2,3,\cdots,M)$,$T^j$是$A$中的第$j$个子集. 为了充分逼近无标签数据的分布,采用的是放回随机采样方法进行子集采样,其中,随机采样子集占比p为20%、40%、70%. 如算法1中getSubset函数所示,将T和子集占比p作为输入,输出的每个T子集大小都符合p要求. 利用放回的随机采样生成的T子集$T^j=\{X_{T^j}^i\}(i=1,2,3,\cdots,H\times p;j=1,2,3,\cdots,M)$,$X_{T^j}^i$是第$j$个$T$子集中第$i$个样本数据.

为了使多个分类器一致拟合数据总体分布,弱标签生成器委员会的联合损失函数定义如图3所示. 图中Y_S是S的真实标签,分类器$h^j(j=1,2,3,\cdots,M)$对输入S分别得到伪标签Y_{f_S}和弱标签Y_{g_S}. 对输入T^j分别得到伪标签$Y_{f_{T^j}}(j=1,2,3,\cdots,M)$和弱标签$Y_{g_{T^j}}(j=1,2,3,\cdots,M)$.

通过上述5个参数得到$h^j(j=1,2,3,\cdots,M)$的损失$Y_{\text{loss}}^j(j=1,2,3,\cdots,M)$,损失函数表示如式(1)所示:

$$\begin{aligned}Y_{\text{loss}}^j=&\text{categorical\_crossentropy}(Y_S,Y_{f_S})\\&+\text{categorical\_crossentropy}(Y_S,Y_{g_S})\\&+\text{categorical\_crossentropy}(Y_{f_{T^j}},Y_{g_{T^j}})\end{aligned}\tag{1}$$

在式(1)中,categorical_crossentropy表示类别交叉

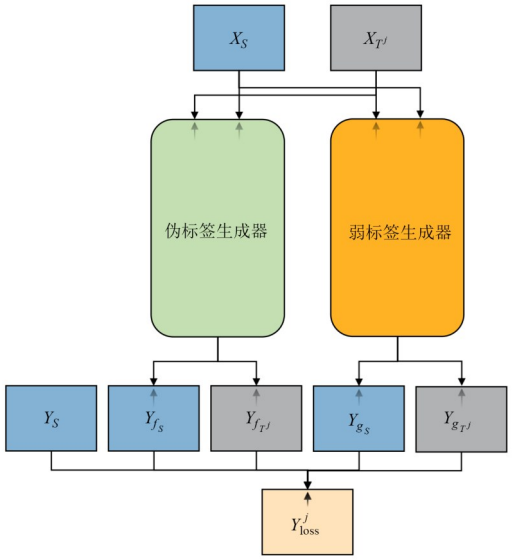

图3 分类器损失

熵损失函数. 对所有分类器损失$Y_{\text{loss}}^j(j=1,2,3,\cdots,M)$进行求和,获得弱标签生成器委员会的联合损失$Y_{\text{loss}}$,如式(2)所示:

$$Y_{\text{loss}}=\sum_{j=1}^{M}Y_{\text{loss}}^j\tag{2}$$

将T作为分类器h_w^M中伪标签生成器f_w^M的输入获得伪标签Y_{f_T}. 利用分类器中的弱标签生成器组成规模为M的委员会,将T依次作为各个弱标签生成器委员$g_w^j(j=1,2,3,\cdots,M)$的输入获得其弱标签$Y_{g_T^j}(j=1,2,3,\cdots,M)$,同时,利用generateConf函数得到$Y_{g_T^j}(j=1,2,3,\cdots,M)$的置信度[17]$C_{g_T^j}(j=1,2,3,\cdots,M)$. 为了计算本文方法中的置信度,将$x_i$作为输入,弱标签生成器的输出结果定义如式(3)所示:

$$\{y_i^1,\cdots,y_i^b,\cdots,y_i^B\}=g_\theta(x_i)\tag{3}$$

其中,y_i^b是指x_i属于b类的概率,如式(4)所示:

$$y_i^b\propto P(y_i=b|x_i)\tag{4}$$

分类结果的置信度定义为如式(5)所示:

$$C=\arg\max_{1\leqslant b\leqslant B}P(y_i=b|x_i)\tag{5}$$

本文利用置信度来体现委员会的可靠性. 受最不确定查询策略[19]的启发,尝试探索委员会预测结果的最低、平均和最高置信度对模型性能的影响. 根据弱标签生成器委员会得到的$Y_{g_T^j}(j=1,2,\cdots,M)$置信度$C_{g_T^j}(j=1,2,\cdots,M)$和可选置信度类型type={min, mean, max},计算出不同类型的置信度C_T,如算法1中getConf函数所示. 当输入的type=min时,getConf函数利用式(6)获得委员会预测结果的最低置信度Min_C_{g_T}. 输入的type=mean或max时,getConf函数利用式(7)和式(8)获得委

员会预测结果的平均置信度 Mean_C_{g_T} 或最高置信度 Max_C_{g_T}. 这里,置信度 $C_{g_T^j}=\left\{C_{g_T^j}^i\right\}(i=1,2,3,\cdots,H\times p;j=1,2,3,\cdots,M)$,$C_{g_T^j}^i$ 是第 j 个弱标签生成器得到的 $Y_{g_T^j}$ 中第 i 个样本的置信度.

$$\text{Min\_}C_{g_T}=\min_{1\leqslant j\leqslant M}\left\{C_{g_T^j}^i\right\}\quad(i=1,2,3,\cdots,H\times p;j=1,2,3,\cdots,M)\tag{6}$$

$$\text{Mean\_}C_{g_T}=\operatorname{mean}_{1\leqslant j\leqslant M}\left\{C_{g_T^j}^i\right\}\quad(i=1,2,3,\cdots,H\times p;j=1,2,3,\cdots,M)\tag{7}$$

$$\text{Max\_}C_{g_T}=\max_{1\leqslant j\leqslant M}\left\{C_{g_T^j}^i\right\}\quad(i=1,2,3,\cdots,H\times p;j=1,2,3,\cdots,M)\tag{8}$$

算法 1　弱标签生成器委员会

输入: 有标签样本数据 X_S, 对应的样本标签 Y_S, 无标签样本数据 X_T, 伪标签生成器 f, 弱标签生成器 g, 分类器 h

输出: T 伪标签 Y_{f_T}, T 弱标签集合 $Y_{g_T^j}$ 及其置信度 C_T

1. $f_w=f(X_S,\ Y_S)$
2. $Y_T'=f_w(X_T)$ //大量无标签数据的伪标签
3. $g_w=g(X_T,\ Y_T')$
4. for j in M do
5. $\quad T^j=\text{getSubset}(T,\ p)$ //获得 T 子集 T^j
6. $\quad h_w^j=h(X_S,\ Y_S,\ X_{T^j})$ //X_{T^j} 是 T 子集 T^j 中的数据
7. $\quad Y_{g_T^j}=g_w^j(X_T)$
8. $\quad C_{g_T^j}=\text{generateConf}(Y_{g_T^j})$
9. end
10. $Y_{f_T}=f_w^M(X_T)$
11. $C_T=\text{getConf}(C_{g_T^1},\cdots,C_{g_T^M},\ \text{type})$

2.4　查询策略

查询策略研究的关键在于如何选择出最有价值的样本,因此,本文的查询策略基于弱标签生成器委员会. 在弱标签生成器委员会中,M 个弱标签生成器委员输出的 T 弱标签 $Y_{g_T^j}(j=1,2,\cdots,M)$ 会出现分歧. 于是,根据少数服从多数原则来决定委员会输出的 T 弱标签 Y_{g_T}. 对于 T 中的每个样本,选择委员会中出现次数最多的标签作为弱标签,通过算法 2 中 getLabelvoted 函数实现. 利用 T 伪标签 Y_{f_T} 和弱标签 Y_{g_T} 筛选出具有分歧的样本,在这些样本中根据弱标签置信度 C_T 进行样本筛选得到无标签样本下标 id_{man},如算法 2 中 getIdman 函数所示. 具体而言,先筛选出 Y_{f_T} 和 Y_{g_T} 不一致的样本下标,也称为候选样本 id_n. 这些样本是模型难以区分的样本,具有更强代表性. 在候选样本中利用 C_T 进行样本筛选,需要对 C_T 进行升序排序获 C_{order}. 利用 C_{order} 筛选出 D 个置信度低于 C_select 的样本下标 id_1,通过求 id_n 和 id_1 的并集获得 id_{nl}. 在实验过程中若出现 id_{nl} 的数据量小于 D 的情况,利用 C_{order} 筛选出置信度低于填充样本置信度标准 s 的样本,并求该样本与 id_{nl} 的并集获得新的 id_{nl}. 若 id_{nl} 的数据量仍小于 D,则利用置信度增量 Δs 不断更新 s,直到筛选出 D 个样本. 最终,筛选出的样本进行人工标注,再利用人工标注后的样本对 S 和 T 进行更新操作.

算法 2　查询策略

输入: T 伪标签 Y_{f_T}, T 弱标签集合 $Y_{g_T^j}$ 及其置信度 C_T, 样本筛选置信度标准 C_select, 填充样本置信度标准 s 和置信度增量 Δs

输出: 无标签样本下标 id_{man}

1. $Y_{g_T}=\text{getLabelvoted}(Y_{g_T^1},Y_{g_T^2},\cdots,\ Y_{g_T^M})$
2. $id_{\text{man}}=\text{getIdman}(Y_{f_T},Y_{g_T},C_T,D)$
3. $X_S=X_S\cup(X_T[\text{id}_{\text{man}}])$
4. $X_T=X_T-(X_S[\text{id}_{\text{man}}])$
5. $Y_S=Y_S\cup Y_{\text{man}}$//$Y_{\text{man}}$ 是人工标注的标签
6. Function getIdman(Y_{f_T},Y_{g_T},C_T,D)
7. $\text{id}_n=\text{where}\,(Y_{f_T}\neq Y_{g_T})$
8. $\text{id}_1=\text{where}\,(C_{\text{order}}[D]\leqslant C\text{\_select})$
9. $\text{id}_{nl}=\text{id}_n\cap\text{id}_1$
10. while $\text{getSize}(\text{id}_{nl})\leqslant D$ do
11. $\quad\text{id}_{nl}=\text{id}_{nl}\cup\text{where}(C_T\leqslant s)$
12. $\quad s+=\Delta s$
13. end while
14. return id_{nl}

3　实验

3.1　实验设置

3.1.1　数据描述

为有效评估本文标注方法效果,在 3 个公开数据集上进行实验. 数据集细节如表 1 所示,表中计量单位为条.

表 1　数据集统计表

数据集	Seed	T	S
IMDB	500	1 136	284
20NEWS	500	1 141	286
chnsenticorp	3 040	1 200	760

IMDB(Internet Movie DataBase):该数据集是基于互联网电影数据库的情节总结分类数据构成,经过筛选后,由 2 个类别构成,分别是动作和浪漫类别. 该数据集包含 1 920 条标注数据,先筛选 500 条样本作为种子集 Seed,其余数据样本按照 8∶2 的比例划分为大量无标签数据 T 和少量有标签数据 S.

20NEWS(20NEWSgroup):该数据集是用于文本分类、文本挖掘和信息检索研究的国际标准数据集之

一,经过处理只包含正文、来源和主题,并且只保留了1 927条数据,数据处理方式同上.

chnsenticorp(chnsenticorp_htl_all):该数据集是由哈尔滨工业大学开源的数据集,该数据集主要用于情感二分类的应用.经过筛选抽取了5 000条标注数据,数据处理方式同上.

3.1.2 评价指标

在本文中,采用以下3种评价指标:F_1分数(F_1-Score)、精确率(Precision)和召回率(Recall).F_1分数综合考虑了精确率和召回率,可以反映整体性能,F_1分数指标越大说明性能更好.F_1分数的如式(9)所示:

$$F_1\text{-Score}=2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{9}$$

精确率如式(10)所示:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

召回率如式(11)所示:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

其中,TP(True Positive)表示真正的正例样本,FN(False Negative)表示假的负例样本,FP(False Positive)表示假的正例样本,TN(True Negative)表示真正的负例样本.

3.1.3 超参数设置

本实验是迭代的半自动标注过程,针对每个数据集分不同置信度和随机子集采样比例做对比实验,置信度分为最小值、最大值、平均值3种情况,子集采样分别选取20%、40%、70%放回随机采样方法.经过多次实验,本文方法的主要参数设置如表2所示.

表2 本文方法的主要参数

参数	值
最大迭代次数	100
T子集数量	5
嵌入层词维度	768
句子最大长度	200
BiGRU	128
Dropout	0.25
优化器	Adam
Batch Size	32

3.2 实验结果与分析

本文提出的标注方法实验性能采用2种评价方式,即数据标注和分类任务的准确性.数据标注的准确性评价是利用种子集Seed作为测试集,将标注结果与真实标签进行比较.分类任务的准确性评价是将人工标注后的T作为训练集,用于下游终端模型的训练并利用种子集Seed作为测试集,评价训练结果.为了选出1个通用的置信度和随机采样子集占比组合的标注模型,将置信度和随机采样子集占比的多种组合在IMDB、20NEWS和chnsenticorp数据集上分别进行性能比较.

3.2.1 数据标注的准确性评价

本文数据集数据标注的准确性评价结果如表3,评价指标依次为F_1分数、精确率Precision和召回率Recall.置信度和随机采样子集占比组合为9种情况,对应于单个数据集的9个实验,表3中加粗部分表示置信度和子集占比组合在某数据集综合表现最好的实验结果.从IMDB数据集的9个实验结果可以看出,在取平均置信度和70%子集占比情况下,F_1分数和召回率最高,精确率表现一般,与最好的精确率相差接近7%.在取最小置信度和40%子集占比情况下,精确率最高,召回率表现最差且F_1分数也偏低.综合分析,IMDB数据集在本文提出的标注方法中,选取平均置信度和70%

表3 本文数据集的数据标注准确性结果

数据集	置信度	子集占比/%	F_1	Precision	Recall
IMDB	最小	20	80.13	81.22	79.06
		40	82.86	93.30	74.52
		70	85.30	85.17	85.44
	平均	20	80.52	82.27	78.84
		40	86.64	88.07	85.25
		70	**88.32**	**86.41**	**90.32**
	最大	20	80.23	75.41	85.71
		40	81.45	86.61	76.87
		70	87.86	90.54	85.35
20NEWS	最小	20	90.93	90.33	91.53
		40	93.50	91.21	95.92
		70	94.58	92.50	96.75
	平均	20	91.89	92.02	91.77
		40	**96.68**	**96.05**	**97.33**
		70	95.25	93.68	96.86
	最大	20	94.88	92.82	97.05
		40	93.66	94.27	93.05
		70	93.61	93.86	93.36
chnsenticorp	最小	20	59.17	54.28	54.28
		40	56.76	54.53	59.18
		70	59.40	54.30	65.55
	平均	20	58.24	54.57	62.44
		40	62.19	54.86	71.79
		70	57.11	56.01	58.24
	最大	20	59.95	54.78	66.19
		40	58.96	54.53	64.17
		70	**62.88**	**53.59**	**76.07**

子集占比组合表现最佳.

与此类似,在20NEWS和chnsenticorp数据集上综合表现最佳的置信度和子集占比组合也很突出.

3.2.2 分类任务的准确性评价

本文数据集分类任务的准确性评价结果如表4,与数据标注的准确性评价方式分析方法相同,表中加粗部分表示置信度和子集占比组合在数据集中综合表现最好的实验结果.

表4 本文数据集的分类任务准确性结果

数据集	置信度	子集占比/%	F_1	Precision	Recall
IMDB	最小	20	77.62	72.86	83.05
		40	77.22	92.85	66.10
		70	72.41	86.47	62.28
	平均	20	78.86	75.78	82.20
		40	67.19	62.96	72.03
		70	**80.08**	**76.24**	**84.32**
	最大	20	73.20	62.38	88.55
		40	79.28	74.16	85.16
		70	79.39	81.33	77.54
20NEWS	最小	20	78.33	81.36	75.52
		40	85.33	77.77	94.51
		70	90.18	85.87	94.93
	平均	20	88.69	89.65	87.76
		40	**93.72**	**92.94**	**94.51**
		70	91.75	87.69	96.20
	最大	20	87.14	79.92	95.78
		40	89.23	83.21	96.20
		70	89.72	89.16	90.29
chnsenticorp	最小	20	68.13	54.30	91.43
		40	57.68	53.36	62.76
		70	63.04	56.80	70.84
	平均	20	60.97	52.10	73.47
		40	57.01	53.29	61.28
		70	**69.79**	**53.84**	**99.17**
	最大	20	59.67	54.40	66.06
		40	60.37	55.35	66.39
		70	67.33	50.79	99.83

3.2.3 与现有方法的性能对比

为验证本文所提出方法的有效性,将本文方法和拥有良好表现的现有文本数据标注方法在3个数据集上进行了整体性能比较. 通过综合数据集在2种评价方式中的结果可知,本文中数据集在取平均置信度和70%子集占比组合情况下,2种评价结果都较好. 该组合模型下关于本文方法与现有标注方法在数据标注上的准确性结果如表5所示.

表5 平均置信度和70%子集占比组合模型的数据标注准确性结果

数据集	方法	F_1	Precision	Recall
IMDB	User Heuristics	32.28	94.84	19.45
	Snuba	58.10	81.09	45.27
	DWLDAL	88.32	86.14	90.32
20NEWS	User Heuristics	55.38	87.92	40.42
	Snuba	81.18	89.13	74.54
	DWLDAL	95.25	93.68	96.86
chnsenticorp	User Heuristics	28.61	47.77	20.42
	Snuba	54.51	55.88	53.21
	DWLDAL	57.11	56.01	58.24

同样,平均置信度和70%子集占比组合模型关于本文与现有标注方法在分类任务上的准确性结果如表6所示. 在2种评价方式中,本文提出的标注方法在IMDB、20NEWS和chnsenticorp数据集上F_1分数与现有方法相比都有显著提升.

表6 平均置信度和70%子集占比模型的分类任务准确性结果

数据集	方法	F_1	Precision	Recall
IMDB	User Heuristics	53.15	59.52	48.01
	Snuba	79.07	78.42	79.74
	DWLDAL	80.08	76.24	84.32
20NEWS	User Heuristics	43.33	32.76	63.97
	Snuba	69.03	53.19	98.31
	DWLDAL	91.75	87.69	96.20
chnsenticorp	User Heuristics	38.95	29.81	56.17
	Snuba	64.96	50.83	89.96
	DWLDAL	69.79	53.84	99.17

从表5可以计算得到,在数据标注评价方式中,DWLDAL在IMDB、20NEWS、chnsenticorp上相较最先进的标注方法Snuba在F_1分数上分别提升了30.22%、14.07%、2.57%. 从表6可以计算得到,在分类任务准确性评价方法中,DWLDAL在IMDB、20NEWS、chnsenticorp上相较Snuba在F_1分数上分别提升了1.01%、22.72%、4.83%. 提升的主要原因有2个:(1)在深度学习模型中,本文设计了弱标签生成器委员会,在捕获数据语义信息和减少单个模型的偏差方面有重要作用. 特别地,采用随机采样的子集进行训练,尽可能拟合了无标签数据总体;(2)不同置信度决策方式对提升模型性能有作用,尤其在查询策略中筛选出最有价值的样本,有效降低了人工标注的成本.

4 结论

本文提出了基于弱标签争议的半自动分类数据标注方法. 为了选择出最有价值的样本,通过基于委员会查询策略筛选出模型难以区分且可信度高的样本. 文章提出一种弱标签生成器委员会模型,可以让查询策略更加便捷地应用在DL上,并充分学习文本信息特征. 本文在3个数据集上进行了大量实验,对比置信度与子集占比组合的多个实验效果,最终得到通用的平均置信度与70%子集占比组合的标注模型. 同时,对比了本文方法和优秀的Snuba方法的自动标注效果在相同数据集中的实验结果,证明本文的方法拥有更高的数据标注准确度. 在未来的工作中,笔者计划在更大的数据集上做深层次分析,争取实现自动标注的基础上减少人工标注量.

参考文献

[1] CAO Z H, WONG K, LIN C T. Weak human preference supervision for deep reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5369-5378.

[2] 何雨航. 基于深度神经网络与弱监督学习的开放域问答技术研究[D]. 兰州: 兰州大学, 2022.
HE Y H. Research on Open Domain Question Answering Technology based on Deep Neural Networks and Weak Supervised Learning[D]. Lanzhou: Lanzhou University, 2022. (in Chinese)

[3] REN P Z, XIAO Y, CHANG X J, et al. A survey of deep active learning[J]. ACM Computing Surveys, 2022, 54(9): 1-40.

[4] TRUST P, ZAHRAN A, MINGHIM R. Understanding the influence of news on society decision making: Application to economic policy uncertainty[J]. Neural Computing and Applications, 2023, 35(20): 14929-14945.

[5] RATNER A, BACH S H, EHRENBERG H, et al. Snorkel: Rapid training data creation with weak supervision[J]. The VLDB Journal, 2020, 29(2): 709-730.

[6] VARMA P, RÉ C. Snuba: Automating weak supervision to label training data[J]. Proceedings of the VLDB Endowment, 2018, 12(3): 223-236.

[7] Park Y, Han D J, Park J W, et al. Distribution aware active learning via gaussian mixtures[C]//The International Conference on Learning Representations. Washington: ICLR, 2023: 1-22.

[8] CHEN Y K, CARROLL R J, HINZ E R M, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data[J]. Journal of the American Medical Informatics Association, 2013, 20(e2): e253-e259.

[9] GOUDJIL M, KOUDIL M, BEDDA M, et al. A novel active learning method using SVM for text classification[J]. International Journal of Automation and Computing, 2018, 15(3): 290-298.

[10] BUCHERT F, NAVAB N, KIM S T. Toward label-efficient neural network training: Diversity-based sampling in semi-supervised active learning[J]. IEEE Access, 2023, 11: 5193-5205.

[11] ZHOU S S, CHEN Q C, WANG X L. Active deep networks for semi-supervised sentiment classification[J]. Coling 2010-23rd International Conference on Computational Linguistics, Proceedings of the Conference, 2010, 2: 1515-1523.

[12] BHATTACHARJEE S D, TALUKDER A, BALANTRAPU B V. Active learning based news veracity detection with feature weighting and deep-shallow fusion[C]// 2017 IEEE International Conference on Big Data. Piscataway: IEEE, 2017: 556-565.

[13] LISON P, BARNES J, HUBIN A. Skweak: Weak supervision made easy for NLP[C]//59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2021: 337-346.

[14] LIU P, WANG L Z, RANJAN R, et al. A survey on active deep learning: From model driven to data driven[J]. ACM Computing Surveys, 2022, 54(10s): 1-34.

[15] LIU J, YANG Y H, LV S Q, et al. Attention-based BiGRU-CNN for Chinese question classification[J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 12(2): 709-730.

[16] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019: 4171-4186.

[17] GAO Z W, LI Z Y, LUO J Y, et al. Short text aspect-based sentiment analysis based on CNN + BiGRU[J]. Applied Sciences, 2022, 12(5): 2707.

[18] SAFRANCHIK E, LUO S Y, BACH S. Weakly supervised sequence tagging from noisy rules[J]. Proceedings

of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5570-5578.

[19] BELUCH W H, GENEWEIN T, NURNBERGER A, et al. The power of ensembles for active learning in image classification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9368-9377.

作者简介

李自强　男,1970年8月出生,四川眉山人. 四川师范大学教授,硕士生导师. 主要研究方向为机器学习、智慧教育、自然语言处理.
E-mail: zqliwww@163.com

杨　薇　女,1999年10月出生,四川南充人. 西南石油大学计算机与软件学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: 1347547762@qq.com

杨先凤　女,1974年5月出生,四川南充人. 西南石油大学教授,硕士生导师. 主要研究方向为计算机图像处理、智慧教育、数据库技术.
E-mail: 565695835@qq.com

罗　林　男,1997年7月出生,四川南充人. 泰豪软件股份有限公司成都研发中心工程师. 主要研究方向为自然语言处理.
E-mail: 1830443303@qq.com