# 技术交底书

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **基本信息** | | | | | | | | |
| **提案名称** | 基于特征插值的两阶段文本主动学习算法 | | | | | | | 提案人觉得合适即可 |
| **提案日期** | 2022.4.23 | | **应用项目** | | 发明√ 新型□ | | | 填表日期，应用范围 |
| **发明人** | 曾杰林 | 李林静 | | 梁嘉琦 | | 曾大军 |  |  |
| **技术 联系人** | 李林静 | | **Email** | | linjing.li@ia.ac.cn | | | 技术联系人将负责整个申请过程中的技术支持 |
| **电话 （座机）** | 010-82544709 | | **电话 （手机）** | | 15810583254 | | |
| **提案详情** | | | | | | | | |
| **一、现有技术（背景技术）** | | | | | | | | |
| 近年来，数据注释已成为深度学习商业化的障碍。为了提升数据标注的效率，现有工作采取主动学习算法（Active Learning, AL），从未标记数据池中选择信息最丰富的实例，供专家和注释员标注。目前已有的主动学习算法，主要是基于候选池样本的预测不确定性和分布多样性进行采样，尽管取得了很大进展，但当前的AL方法在应用于高维数据和缺乏数据的领域时仍然很困难，近期研究提出了基于特征插值的主动学习算法，其通过已训练模型的特征提取模块，计算每个类别已标注样本的均值特征，随后将未标注样本与上述计算的各个类别均值特征进行插值，从而得到新特征，最后比对插值前后特征对应的分类结果，决定是否选择该样本进行标注（Parvaneh A, Abbasnejad E, Teney D, et al. Active Learning by Feature Mixing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12237-12246.）。该算法核心在于通过将未标注样本的特征向量与已标注样本的特征向量进行插值，来发现模型未曾见过或目前无法识别的特征（后续称为“新特征”），并选择包含此类新特征的未标注样本。但我们发现该算法虽然能发现包含新特征的样本，但某些新特征对于模型更新的贡献很小，选择该类样本进行标注对于模型性能的提升不大，其并未考虑到此种问题，导致其算法在几个文本基准数据集上的表现不佳。 | | | | | | | | |
| **二、本方案创造的目的（本申请解决的技术问题）** | | | | | | | | |
| 为解决现有基于特征插值的主动学习算法无法判断新特征对于分类器是否关键的技术挑战，本项研究使用两阶段主动学习算法：我们认为模型作出错误决策的原因是因为当前样本中包含了模型无法识别的特征，所以在第一阶段利用特征插值技术来发现模型无法识别的特征，并将包含该类特征的未标注样本加入候选池中；在第二阶段使用K近邻搜索技术，逐个判断候选池中样本所包含的特征对模型分类是否关键。利用两阶段的过程，筛选出模型无法识别且对分类有决定性作用的特征，并选择包含该类特征的样本进行标注，提升该类主动学习算法在文本基准数据集上的表现。 | | | | | | | | |
| **三、本方案的技术内容（技术方案）** | | | | | | | | |
| 本方案的技术方案如下所示。  **算法描述：**本方案提出了一种基于特征插值的两阶段主动学习算法，整个算法是一个多轮次的迭代过程，每一轮次的输入数据为大规模未标注的文本以及少数已标注文本，输出结果是算法筛选出来的固定数量该轮次待标注文本。其中核心算法包括1）采用特征插值方法发现模型未曾学习到的特征、2）采用K近邻搜索技术判断新特征的重要性。  **算法流程：**  1）**使用特征插值方法发现新特征**，首先引入预训练语言模型Bert作为文本的编码器，对所有已标注文本提取特征向量，分别计算每个类别所有文本特征向量的均值，作为该类别的语义向量；然后针对每个未标注的候选文本，使用特征插值的方法，将候选文本的特征向量与各个类别的语义向量分别进行线性插值，并对插值的结果做分类结果预测，如果与各个类别插值后的预测结果不一致，则认为该候选文本很可能包含模型无法识别的特征，模型无法对其作出肯定的判断，从而其会因为轻微的插值而改变预测结果。  2）**采用K近邻搜索判断新特征的重要性**，经过第一阶段的筛选，候选池中的文本均包含模型无法识别的特征，但因为每一轮次可进行标注的文本数量有限，因此需要对候选池进行进一步筛选。我们假设特征空间中距离足够近的文本会包含类似的特征，如果该特征对分类结果具有决定性的贡献，例如“梅西”这一特征可足够判断该段文本属于体育类别，那么包含该特征的文本分类标签就应该相近，标签分布的熵值应该很小。因此，针对候选池中的每个未标注文本，使用K近邻搜索技术，寻找特征空间中欧氏距离最近的k个已标注文本，然后收集这k个已标注文本标签，计算其分布的熵值。我们认为计算得到熵值低的文本包含模型无法识别且对分类结果有决定性作用的特征。按照熵值从低到高进行排序，最终选择位于前列固定数量的文本作为这一轮次应该标注的文本。 | | | | | | | | |
| **四、本方案的附图** | | | | | | | | |
| 本方案的技术方案如下图1所示：    图1 方法一个迭代轮次的流程图 | | | | | | | | |
| **五、具体实施方式** | | | | | | | | |
| 根据上述技术方案，本部分将详细描述实施方式。  主动学习算法的语料库是大量的未标注文本和少量的已标注文本，在每一个迭代轮次里，首先根据当前模型和算法从未标注文本中筛选出固定数量的文本，并进行标注，再加入已标注文本池，最后利用已标注的文本对模型进行训练和更新，完成一个迭代轮次。如此迭代直至模型的预测准确率不再提升或达到预期设定目标。  **1) 使用特征插值方法选取新特征**  在每一个迭代轮次开始的时候，使用当前的Bert/Roberta等预训练语言模型提取所有已标注文本向量表示，并选择作为文本的向量表示，然后分别计算每个类别所有文本特征向量的均值，作为该类别的语义向量。  其中是类别的语义向量，是已标注文本中属于类别的样本数量。针对每个未标注的候选文本，同样使用当前模型提取文本向量表示，将候选文本的特征向量与各个类别的语义向量分别进行线性插值，得到各个类别的插值结果，然后利用当前迭代轮次训练状态的线性分类层对插值向量进行分类预测。  如果各个插值向量的预测结果不一样，则说明该候选文本很可能包含模型无法识别的特征，模型无法对其作出肯定的判断，从而会因为轻微的插值而改变预测结果。基于此种推理，我们认为此类未标注文本包含了模型目前无法识别的特征，即新特征。将其全部放入候选池中以待第二阶段的进一步筛选。  **b) 采用K近邻搜索判断特征重要性**  在经过第一阶段筛选后，候选池中的每个未标注文本都包含模型目前无法识别的特征，然而我们发现其中的许多特征对于分类并没有贡献，例如一些生僻字词，且前期的实验过程中发现候选池的规模往往大于该轮次可标注样本的数量，因此在第二阶段，我们对这些新特征的重要性进行排序，并选择对模型分类有决定意义的特征。  通常来讲，如果一个新特征对于模型进行预测至关重要，那么包含该特征的实例将具有相同的标签。基于该假设，我们提出了第二阶段的特征重要性排序算法：针对候选池中的每个文本，提取特征向量后，使用KNN算法在特征空间中搜索到当前最近的k个最近邻已标注文本，并构建该候选样本的空间邻域集：  其中是已标注的样本池，是特征空间中的欧式距离，是一个提前设定的极小超参数。得到样本的邻域集之后，我们针对邻域集中标签的分布，计算熵值：  在依次对候选池中的所有候选文本进行计算之后，我们按熵值的倒序排列所有的样本，并选择排在前列固定数量的文本，作为这一迭代轮次最终确定的要进行标注的文本。  上述便是一个筛选文本的过程，整个算法将按照上述的筛选——标注——模型训练及更新的流程进行循环，直至算法预测的准确率达到了预期目标或者预测的准确率不再提升。 | | | | | | | | |
| **六、本方案的优点（技术效果）**  介绍本方案的优点或进步效果，各优点最好与本方案内容一一对应。 | | | | | | | | |
| 本方案的优点主要如下：  第一，本方案从文本包含的特征入手，探究新的主动学习筛选思路。使用特征插值的方法，发现模型当前无法识别的特征，保证标注人员可以优先标注包含当前模型无法识别的特征的文本。该方法可以极大地提高标注数据的多样性，保证模型在数据标注的过程中迅速地提升性能。  第二，本方案提出了基于特征的两阶段主动学习算法，在提高标注数据多样性的同时，于第二阶段使用KNN技术分析邻域的标签分布，确保标注数据中特征的关键性。进一步提升了数据筛选的合理性，在标注数据规模逐渐增大的情况下可稳定地提升模型性能。  综上所述，本方案可以在主动学习的框架下，在文本的特征层面筛选数据进行标注，且两阶段的筛选过程保证了标注数据中的特征，既是模型当前无法识别的，也是对分类有决定性意义的。在标注数据规模逐渐增大的情况下，具有模型性能提升更快，更稳定的优点。  为验证上述优点，我们将目前最优的几种文本主动学习算法和本方案中提出的算法在SST-2文本情感分类数据集上做了实验验证，考虑到主动学习具有一定的随机性特点，我们对每种算法使用不同的随机数种子运行了5次实验，并对结果取平均值，最终的实验结果如下图：    足以证明本方案的算法（红线）在标注数据规模逐渐增大的情况下，具有模型性能提升更快，更稳定的优点。 | | | | | | | | |
| **七、拓展方案**  本方案有无其它拓展方案或替换方案。 | | | | | | | | |
| 无 | | | | | | | | |