# 技术交底书

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **基本信息** | | | | | | | | |
| **提案名称** | 基于特征插值的两阶段主动学习算法 | | | | | | | 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** | 李林静 | 曾杰林 | | 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与已标注样本的特征进行插值，来发现包含明显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，其并未考虑到此种问题，导致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）** | | | | | | | | |
| 为解决现有聚类算法无法有效理解中文ASR转录文本语义的技术挑战，本项研究使用平均池化方法提取句子的向量表示，同时将聚类中心点作为该类的语义表示，将样本与其所属类别的中心作为一对正样本，与其他类别的中心作为一对负样本，构建语义级别的对比损失。提升聚类算法在中文ASR转录文本聚类任务上的表现。 | | | | | | | | |
| **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对比学习的中文ASR转录文本聚类算法，输入数据为大规模未标注的中文转录文本，输出结果为转录文本的聚类结果。其中核心算法包括1）采取均值池化方法提取文本表示、2）构建语义级别的对比损失。  **算法流程：**  1）**加入均值池化层提取文本表示**，引入中文预训练语言模型Bert作为文本的编码器，并在Bert的输出层上加入一个均值池化层，得到一个固定维度的句子向量表示，减轻常见错误字对语义的影响，保留文本的语义信息。方便后续进行聚类。  2）**构建语义级别的对比损失**，将对比学习的思想引入聚类任务，为无标注中文文本创建正负样本对，提出了文本层面和文本类别层面的两种损失函数，在聚类过程中捕捉文本的高阶语义信息。具体来讲，新的对比损失包含两个损失项，第一项是文本层面的对比损失，希望样本与基于自身生成的正样本在特征空间中更近，与其他样本在特征空间中更远；第二项是类别层面的对比损失，希望样本与自身所属类别的聚类中心在特征空间中更近，与其他类别聚类中心在特征空间中更远。依照此损失函数迭代式地进行训练，直到达到指定的迭代轮次或者聚类的分类准确率达到了预期目标。 | | | | | | | | |
| **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。  **1) 加入均值池化方法提取文本表示**  聚类任务的语料库是大量的中文ASR转录文本，当使用Bert/Roberta等预训练语言模型提取文本向量表示的时候，一段转录文本会被编码为一个向量集合，集合中的每个向量表示文本中的字经过编码之后的字向量，通常情况下，使用作为文本的向量表示，但是在聚类任务中，因为下游不会连接任何分类层，聚类时无法有效捕捉语义信息，所以本方案放弃使用。对文本编码后向量集合中的所有向量作均值处理，提取文本的向量表示。  其中，为句子的字符数，为句子起始标识符的编码向量，为句子结束标识符的编码向量。因为文本的语义信息，例如情感倾向，通常是由文本中大部分字符的情感倾向共同决定的，所以本方案使用均值池化方法来提取文本的向量表示，使语料库中具有相似语义内容的文本可以在向量空间中有着更相似的文本表示。此外，均值池化层可以有效减轻文本中少量的错误字词对分类结果的影响。  **b) 构建语义级别的对比损失**  给定语料库中所有转录文本在语义空间中的向量表示，本方案遵循迭代式的训练方法，不断循环地进行聚类和更新编码器。首先定义语义表示空间为一个维实数值向量空间，空间中包含着语料库中个文本经编码器编码后的向量表示。在每一个轮次的开始，使用k均值聚类算法直接对语义空间中的个文本向量进行聚类，得到聚类结果。  为了获取每个类别的语义向量，对聚类结果中的每个类别，分别计算该类的类别中心和聚类松散程度:  其中为聚类结果中属于该类的文本数量，为平滑系数，防止聚类松散程度趋于正无穷。本方案使用类别中心作为该类别的语义向量表示，利用聚类松散程度衡量该类别语义向量的真实性，具体来说，聚类松散程度越小，则说明该类别的文本在空间中分布越集中，越认为可以很好地代表该类别的语义。  在完成聚类和聚类结果分析之后，本方案分别构建了文本层面和类别层面两个对比损失，在文本层面，首先对文本进行了数据增强，具体来说，利用深度学习模型dropout机制的随机性，对同一个文本再次进行编码，得到一个文本第二个向量表示，并将其与之前聚类时所用的向量表示作为一个正样本对，将一个训练批次中两个不同文本的向量表示作为一个负样本对，构建了文本层面的对比损失：  其中表示一个正样本对在空间中的余弦相似度，表示一个负样本对在空间中的余弦相似度，该损失函数的目标是希望样本与基于自身生成的正样本在特征空间中更近，与其他样本在特征空间中更远。  在类别层面，将文本向量与其当前聚类所属类别的类别中心作为一对正样本对，将文本向量与其他类别的类别中心作为一对负样本对，构建了文本层面的对比损失：  其中代表着当前类别的聚类松散程度，上式中聚类松散程度越小，当前类别语义向量提供的损失也就越多。该损失函数的目标是希望样本与其聚类所属类的语义向量在特征空间中更近，与其他类别的语义向量在特征空间中更远，从而使算法更完美地完成按语义对文本进行聚类的目标。  最终损失函数公式为：  其中是平衡两个对比损失的参数。最后使用Adam（一种反向传播算法的变体）来优化损失函数，训练用于获取文本表示的编码器，以获得可训练参数的值。  上述便是一个循环的过程，整个算法将按照上述的提取表示——聚类——编码器更新的流程进行循环，直至算法完成了指定的循环轮次或者聚类的准确率达到了预期目标。 | | | | | | | | |
| **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，本方案对中文ASR转录文本表示的提取方法做出了探究，并通过实验验证了预训练语言模型+平均池化方法是最适用于中文转录文本聚类任务的文本表示提取方法，该方法有效减轻了转录文本中错误字词对分类结果的影响，使语料库中具有相似语义内容的文本在向量空间中有着更相似的文本表示。  第二，本方案提出了中文语义级别的对比损失，在文本层面的对比损失基础上，加入了语义级别的对比损失。在计算损失时充分考虑中文文本蕴含的高阶语义，提升对比学习方法在中文转录文本无监督聚类任务上的表现。  综上所述，本方案可以在无监督的训练模式下，在语义层面优化中文转录文本的文本表示，提升聚类算法在中文转录文本分类任务上的准确率。具有文本分类更准确、应用前景大等优点。 | | | | | | | | |
| **七、拓展方案**  本方案有无其它拓展方案或替换方案。 | | | | | | | | |
| 无 | | | | | | | | |