各位评委，各位老师，大家好。我们是“痛！太痛了”团队，很荣幸我们的方案可以获得认可，参加此次决赛，我是团队的队长，答辩人李一鸣。

我们团队由4名中国科学院大学2022级研究生新生组成，其中3人来自中国科学院计算技术研究所，1人来自中国科学院自动化研究所。团队的研究兴趣涉及半监督学习、智能问答、生物信息学等领域。

下面我将从赛题背景、任务描述、解决方案和实验结果四个方面对我们团队的方案进行介绍

近年来，随着我国科技的进步和创新能力的提升，我国的专利申请量快速增长，专利检索、查新、管理等需求也不断增加。为了满足多元化专利检索需求，提升专利服务质量，通常需要建立多个维度的专利分类体系。但是这些分类体系一般都比较复杂，专业性强，对非专业人员而言使用有一定的困难。

近年来，随着大规模预训练模型的发展，自然语言处理技术在工业界逐渐落地，这也为解决专利分类问题带来了新的思路。相较于传统的文本，专利文本的标注的数据量较为稀少，因此，小样本学习在专利文本分类中至关重要。为了解决上述问题，我们提出了一种基于类别动态阈值伪标签与尾部类别数据增强的小样本文本分类方法。后面将对我们的方法进行详细的介绍。

首先简单介绍一下本赛题的任务。本赛题要求对专利文本数据进行分类，训练集包括958条专利数据，每一条样本包括专利权人、专利标题、专利摘要、分类标签等属性，其中标签经过脱敏处理，共36类。赛题采用Macro-F1指标对于模型性能进行评估，具体定义如上方的公式所示，其中，P\_i 是第 i 个类别的 Precision，R\_i 是第 i 个类别的Recall，n 是类别总数。对于某一个类别 i 来说，Precision 和 Recall 的计算方式如下方的两个公式所示：其中，TP\_i 为类别 i 的真正例，FP\_i 为类别 i 的假正例，FN\_i 为类别 i 的假负例。

我们团队解决方案的整体流程如这张图片所示：首先根据模型在A榜测试集上输出的softmax分数为不同类别的样本设置不同阈值，来生成伪标签扩充训练集规模，并针对尾部样本进行数据增强，缓解了尾部类别的数据稀疏问题。接着，使用扩充后的训练集对ernie-3.0-base语言模型在文本分类任务上进行微调。其中我们将包含不同粒度信息的title、assignee、abstract字段重新组织为一段文本，也是受近年在自然语言处理领域中新兴的提示学习启发。此外，为了提升模型的鲁棒性以及解决类别不平衡问题，我们进一步采用了对抗训练，提升模型的鲁棒性，实现了更加精准的分类。

对抗训练旨在提升模型的鲁棒性，同时也能提高模型的泛化能力，最终实现更加精准的分类，其基本思想是对原始输入样本 x 添加扰动 r\_adv，得到对抗样本 x\_adv，再将对抗样本输入模型训练。我们具体采用了FGM攻击算法，其所施加的扰动与样本的梯度方向相同，如公式所示：

我们观察发现训练数据集存在标签不平衡的问题，且少量类别占据了绝大多数的样本，大量的类别却仅有少量的样本，也就是数据的长尾问题。研究指出在长尾识别任务下，尾部类别虽然召回率较低，但却有着较高的精确度，而头部类别则相反。这启发我们根据标签分布情况及模型预测情况为每一个类别设置不同的阈值：头部类别较多，模型学习情况较好，应为其设置较高的阈值，保证其精确度；尾部类别较少，模型学习情况较差，应为其设置较低的阈值，从而获得较高的召回率。

我们为本任务设计了一种类别相关的动态阈值伪标签产生策略，对于每一类别，我们先得到预测标签为该类的所有记录，之后将这些记录按照对于该类预测概率由大到小进行排序，随后我们求其 α 分位数（即前 α %大的预测概率）作为该类别的阈值，若求得的阈值小于一个预先给定的固定阈值 γ，则将其重新置为 γ。利用 α 分位数筛选伪标签可以充分地考虑到各个类别的学习情况，同时固定阈值 γ 作为下限，使得尾部类别的筛选不至于混入过多的噪声样本。具体的算法流程如右表所示。

这种伪标签产生流程可以迭代进行，即训练得到一个较好的模型A，利用其产生的伪标签再训练一个模型B，再利用模型B对无标签测试数据打伪标签，并由此再训练一个新的模型C。我们在比赛中的具体迭代过程如这张图所示。最后我们分别采用线上F1分数为0.600, 0.599, 0.632, 0.642, 0.646的模型为A榜测试数据标记伪标签，并将其并入初始训练集中，分别得到伪标签扩展数据集train\_0，train\_1，train\_2，train\_3，train\_4，并在最终参赛系统的训练中与尾部数据增强策略配合使用。

为了缓解类别不平衡的问题，我们对尾部类别应用了两类数据增强方法。

1. 回译：我们利用谷歌翻译接口，将中文翻译为英法德日韩5种语言，再翻译回中文，得到新的样本。2. 简单文本数据增强，分别是同义词替换、随机插入、随机替换与随机删除。数据增强前的多数类别和少数类别的标签分布如右图所示。可以看出数据增强策略在一定程度上缓解了标签的不平衡问题。

下面展示一下我们团队的实验结果。

通过前期大量的实验，并参考线上提交的F1分数，我们最终选择了效果较好的9个模型作为参赛系统的最终结果。9个模型的具体参数细节各不相同，训练数据也采用了不同阶段得到的伪标签，且大多是在经过多折交叉验证后选取的效果较好的某一折或某几折。同时为了满足模型的总大小不超过2G的要求，我们将单精度模型转换为半精度，实验表明这种转换对于最终的推理结果影响不大。最后我们通过投票法将9个模型集成在一起得到最终的推理结果。最终，集成模型在初赛A榜测试集上达到0.6516的F1分数，在B榜测试集上达到了0.5955的F1分数。

为了验证单独某一模块对最终分类效果的影响，我们在前期实验的过程中进行了一些消融实验，分别对预训练模型、伪标签、训练策略、尾部数据增强和预测头架构进行单独验证。

对于预训练模型来说，目前代表性的开源中文预训练模型有很多种。我们对搜集到的多种预训练模型进行了对比实验，最终的实验结果表明ERNIE 3.0更适用于我们的小样本分类学习任务。

我们通过对动态阈值伪标签技术涉及到的两个超参数 α 及 γ 进行实验，其中，实验编号1为算法的默认参数设置，实验编号2的设置为传统伪标签产生算法采取的固定阈值策略，实验编号3为不采用固定阈值 γ 对于阈值下限进行限制。从实验结果来看，本文提出的伪标签产生策略具有良好的扩展性，可以更好地对抗长尾分布现象，并充分挖掘未标注数据所蕴含的信息。

除FGM这种对抗学习策略之外，我们也尝试了多种其他的对抗训练策略，同时对损失函数进行重加权等尝试，以期更好地适应本次任务特性。最终的实验结果表明FGM与PGD更适用于本次的小样本分类学习任务，然而由于PGD的迭代训练开销较大，我们只将其应用于伪标签生成阶段，在实验阶段我们放弃了对其进行进一步的探索。

针对类别不平衡的问题，我们采用了回译和简单文本数据增强两种方法进行缓解。实验结果表明，数据增强方法对于模型的预测效果有显著提升。

Bert相关预训练模型的输出大致有两种，其中包括模型各层输出的隐藏状态和序列的第一个token的最后一层的隐藏状态，也就是pooler output。其中模型最后一层输出的隐藏状态通常用于命名实体识别，而pooler output的输出通常用于句子分类。同时也有研究表明BERT编码了丰富的语言学层次信息。表层信息特征在底层网络，句法信息特征在中间层网络，语义信息特征在高层网络。如果将模型的前几层和后几层进行平均池化，分类效果会更好。因此我们对不同的预测头架构进行了实验，实验结果表明目前使用的最后一层输出的隐藏状态+多层感知机的分类效果最好。

感谢中国科学院大学高级人工智能课程介绍本次竞赛并鼓励我们参加，感谢CCF BDCI组委会和智慧芽公司的全体相关工作人员。