案情文本要素提取系统

科研导师：姚新、袁博

小组成员：叶梓元、蓝文兴、阮沛钧

课题背景

当前，案多人少已是人民法院工作不争的事实，有人形象地提出“诉讼爆炸”的概念。截至2014年底，全国法院未结案件185. 4万件，同比上升46%，数量同比2013年增加了58.4万件。各省市区法院未结案数量均同比上升[1]。为了解决这一问题，人们开始思考如何用人工智能来帮助法院工作人员更高效地工作。当前人工智能方案主要针对于法院辅助判决，判决书文本分类，判决书文本相似度度量，案情要素提取等方面。对于这个课题，我们针对于解决案情要素提取这一问题，即主要解决如何从大量的法院文本中快速地得到文本特征——一个对大多数法院法官和判决人员来说是很难的问题和挑战。

课题意义

本课题的主要目的是为了将案件描述中重要事实描述从复杂无规律的案情中自动抽取出来，并根据领域专家设计的案情要素体系进行分类。与传统的人工提取做对比，这种要素的自动抽取方案可以大大提升案件语义要素提取的效率，降低人力抽取成本。案情要素抽取的结果可以用于案情摘要、可解释性的类案推送以及相关知识推荐等司法领域的实际业务需求中，并且还可以提高办案效率。

问题定义

为了用数学模型来定义这个问题，我们首先定义了如下符号：

$x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\} \in X, X \in R^d$: x_i表示判决书文本中重要事实描述句子的特征向量，X表示d维的样本空间，即句子特征向量构成的空间

$L = l_1, l_2, \ldots, ln$:表示含有n个标签的标签空间

$|A|$:表示集合A的长度

$D = \{(x_i, Y_i)|1 \leq i \leq |X|, x_i \in X, Y_i \subset L\}$:表示多标签训练集

问题的数学模型为：经过多标签机器学习算法得到一个函数$f : x_i \to Y_i, Y_i \subset L$，能够为每一个样本$x_i$预测其相关的标签集合$Y_i$。

文献综述

由问题定义可知，该问题是一个多标签分类问题，因此我们对多标签的相关文献进行了调研。特别地，计算机无法处理直接的文本，需要将文本转化为计算机能够处理的形式，而目前最流行的是将句子转化为向量，所以我们也对句子向量化进行了调研。同时，我们在处理数据时发现标签数据中正负样本比例十分不平衡，负样本比例显著高于正样本，所以我们还对类别不平衡的相关文献进行了调研。这里本报告主要从文本向量化、多标签分类算法以及类别不平衡/小样本问题三个方面对已有的一些文献进行简要叙述。

1. 文本向量化

文本向量化主要是将文本转化为计算机可以直接处理的向量形式，文本转化成向量后，对应的向量若包含越多原始文本信息，则该向量对文本的表示能力越强。

1.1 Word2Vec

Word2Vec是可以将一个单词转换为矢量的基本算法，用于通过其矢量比较两个单词的相似性。 它有两个模型，分别是连续词袋模型（CBOW）和Skip-gram模型。 这两个模型是相反的，因为CBOW模型使用上下文来预测当前单词，而Skip-gram模型使用当前单词来预测上下文[2]。 Skip-gram模型试图根据同一句子中的其他单词来最大化一个单词应出现的概率[3]。 接下来，我们简要介绍一下Skip-gram模型。 在训练了Skip-gram模型之后，我们想要获得隐藏层中的加权矩阵。 此外，跳过语法模型的输入是当前单词的独热码，而输出是周围单词的矢量。

1.2 Doc2vec

Le和Mikolov [4]提出了doc2vec，作为对word2vec的简单扩展，将嵌入的学习从单个单词扩展到连续单词。与word2vec相似，doc2vec以两种形式提出：DBOW（单词分布袋）和PV-DM（段向量分布内存模型）[5]。DBOW与Skip-gram的工作方式相同，只是输入被段落标记代替，并且忽略了单词的顺序。 PV-DM的工作方式与CBOW相同。 除了多个目标词外，它还引入了一个附加的段落标记作为输入。与CBOW不同，向量是级联而不是求和的。因此，它具有考虑词序的优点，其目的是在给定串联的段落和单词向量的情况下预测上下文单词。

1.3 Bidirectional Encoder Representation from Transformers (BERT)

BERT是截至2018年10月代表最新技术的模型，通过预训练模型和微调模型使其使用于不同的任务的方式，横扫了11项NLP任务。BERT的模型架构是多层双向转换编码器，基于Vaswani等人描述的原始实现，主要包括预训练和微调两部分[6]。在预训练部分包括预测被遮蔽单词和下一句子，前者主要为了训练深度双向表征，通过随机选取15%的单词，而在这些单词中10%的单词会被替换成其他的单词，而10%的单词不会被替换，剩下80%的单词才被用[MASK]来遮蔽，后者可以训练一个理解句子关系的模型，我们预训练了一个二值化下一句预测任务，该任务可以从任意单语语料库中轻松生成[6]。具体来说，选择句子A和B作为预训练样本：B有50% 的可能是A的下一句，也有 50% 的可能是来自语料库的随机句子。在微调部分，对于每个任务，我们只需将特定于任务的输入和输出插入BERT，并端对端微调所有参数[6]。在输入时，来自预训练的句子A和句子B类似于释义中的句子对，含假设的前提对，问答中的疑问句对和一个在文本分类或序列标记中退化文本对[6]。在输出处，将令牌表示形式输入到输出层中，以进行令牌级任务，例如序列标记或问题回答，将表示形式输入到输出层中，以进行分类，例如内容或情感分析[6]。

2. 多标签分类算法

多标签学习的主要难点在于输出空间的爆炸增长，比如20个标签，输出空间就有2^{20}，为了应对指数复杂度的标签空间，需要挖掘标签之间的相关性。比方说，一个图像被标注的标签有热带雨林和足球，那么它具有巴西标签的可能性就很高。一个文档被标注为娱乐标签，它就不太可能和政治相关。有效的挖掘标签之间的相关性，是多标签学习成功的关键[7]。根据对相关性挖掘的强弱，可以把多标签算法分为三类。

- 一阶策略：忽略和其它标签的相关性，比如把多标签分解成多个独立的二分类问题[7]。
- 二阶策略：考虑标签之间的成对关联，比如为相关标签和不相关标签排序[7]。
- 高阶策略：考虑多个标签之间的关联，比如对每个标签考虑所有其它标签的影响[7]。

而多标签学习算法可分为如下两类：问题转换的方法和算法改编的方法。

2.1 基于转化的算法

基于转化的多标签算法旨在将多标签问题转化为其他更易于处理的子问题。以下是对其中几种基本算法的简单介绍:

- 二元关联(Binary Relevance)算法[8]通过将一个在标签空间为$L = l_1, l_2, \ldots, l_n$的多标签问题分解为n个二分类类问题，对每个标签训练一个二分类器，最后利用训练得到的n个分类器预测得到一个样本的标签$L' \subset L$。该方法是建立在n个标签相互独立的假设上的。

- 分类器链(Classifier Chain)算法[9]是将一个多分类问题转化为一个链式的二分类问题，第i个标签$(2 \leq i \leq n)$的预测建立在前$i-1$个标签的预测结果上，通过将前$i-1$个标签的预测结果和样本x输入到第i个分类器，输出样本x的第i个标签的预测结果。
- 随机k标签PowerSet(Random k-Labelsets)算法[10]是利用了Label Powerset算法的思想，将多标签问题转化为一个多分类问题。LP算法是只将一个n维的预测空间映射到一个整数空间，来实现多标签向多分类的问题转化。对于n维的标签空间，利用函数$h:2^n \rightarrow N$，把结果映射为一个整数，通过学习样本和映射后整数的关系来训练一个多分类器。
- Random k-Labelsets算法则对LP算法进行了改进，随机选择t组长度为k的标签序列，训练t个多分类器，最后利用t个预测结果对所有标签进行投票，这样就解决了LP算法在标签空间过大时分类准确率低的问题。

2.2 基于改编的算法

算法改编的方法：通过改编流行的学习算法去直接处理多标签数据，比如改编懒惰学习，决策树，核技巧。以下是对其中几种基本算法的简单介绍:

- *Multi-Label k-Nearest Neighbor(ML-kNN)*: 该算法的基本思想是使k最近邻居技术适应多标签数据，其中使用最大后验规则来进行预测，最大后验规则是通过包含在邻居中的标签信息来推理是否含有标签[7]。
- *Multi-Label Decision Tree (ML-DT)*: 该算法的基本思想是采用决策树技术处理多标签数据，利用基于多标签熵的信息增益准则递归建立决策树[7]。
- *Ranking Support Vector Machine (Rank-SVM)*: 该算法的基本思想是采用最大间隔策略来处理多标签数据，其中优化了一组线性分类器以最小化经验排名损失，并能够处理带有核技巧的非线性情况[7]。
- *Collective Multi-Label Classifier (CML)*: 该算法的基本思想是采用最大熵原理来处理多标签数据，其中标签之间的相关性被编码为结果分布必须满足的约束[7]。

TABLE 2
Summary of Representative Multi-Label Learning Algorithms Being Reviewed

Algorithm	Basic Idea	Order of Correlations	Complexity [Train/Test]	Tested Domains	Optimized Metric
Binary Relevance [5]	Fit multi-label data to q binary classifiers	first-order	$\mathcal{O}(q \cdot \mathcal{F}_{\mathcal{B}}(m,d))/$ $\mathcal{O}(q \cdot \mathcal{F}'_{\mathcal{B}}(d))$	image	classification (hamming loss)
Classifier Chains [72]	Fit multi-label data to a chain of binary classifiers	high-order	$\mathcal{O}(q \cdot \mathcal{F}_{\mathcal{B}}(m,d+q))/$ $\mathcal{O}(q \cdot \mathcal{F}'_{\mathcal{B}}(d+q))$	image, video text, biology	classification (hamming loss)
Calibrated Label Ranking [30]	Fit multi-label data to $\frac{q(q+1)}{2}$ binary classifiers	second-order	$\mathcal{O}(q^2 \cdot \mathcal{F}_{\mathcal{B}}(m,d))/$ $\mathcal{O}(q^2 \cdot \mathcal{F}'_{\mathcal{B}}(d))$	image, text biology	Ranking (ranking loss)
Random k-Labelsets [94]	Fit multi-label data to n multi-class classifiers	high-order	$\mathcal{O}(n \cdot \mathcal{F}_{\mathcal{M}}(m,d,2^k))/$ $\mathcal{O}(n \cdot \mathcal{F}'_{\mathcal{M}}(d,2^k))$	image, text biology	classification (subset accuracy)
ML-kNN [108]	Fit k-nearest neighbor to multi-label data	first-order	$\mathcal{O}(m^2d+qmk)/$ $\mathcal{O}(md+qk)$	image, text biology	classification (hamming loss)
ML-DT [16]	Fit decision tree to multi-label data	first-order	$\mathcal{O}(mdq)/\mathcal{O}(mq)$	biology	classification (hamming loss)
Rank-SVM [27]	Fit kernel learning to multi-label data	second-order	$\mathcal{O}(\mathcal{F}_{\mathrm{QP}}(dq+mq^2,mq^2)$ $+q^2(q+m))/\mathcal{O}(dq)$	biology	Ranking (ranking loss)
CML [33]	Fit conditional random field to multi-label data	second-order	$\mathcal{O}(\mathcal{F}_{\mathrm{UNC}}(dq+q^2,m))/$ $\mathcal{O}((dq+q^2)\cdot 2^q)$	text	classification (subset accuracy)

Figure 1. 多标签学习算法对比表[7]

3. 类别不平衡/小样本学习

类别不平衡主要是从三种角度对算法进行考量。分别是事前补偿，事中补偿和事后补偿。

- 事前补偿主要是在于采样技术的选择上，采样技术又主要分为两种：随机采样和人工采样。
    - 随机采样是最简单也是应用最为广泛的一类采样技术，但是它缺少一些先验分布知识。

    - 人工采样则能弥补随机采样的这个缺陷，通过人工干预的方式添加或移除样本。
- 事中补偿主要是代价敏感学习技术。该技术的主要思想在于：在训练分类模型时，不再以样本的整体误差最小化为训练目标，而是转而追求整体误差划分代价的最小化[11]。
- 事后补偿技术——决策输出补偿技术则是通过对分类器的决策输出值进行调整，以达到修正偏倚分类面的目的[12]。

提出的方法

1. 句子向量化

对于句子向量化这一块我们想使用Google的BERT模型，其原因如下：

- BERT模型是目前自然语言处理 (NLP)中代表最新技术的模型，在其他NLP任务中具有良好的表现。
- 之前我们曾经使用过Doc2Vec模型，但是效果并不是很好。
- 通过这次创新实验，我们想学点新知识，因此想尝试新的模型。

2. 多标签模型

在多标签训练算法上，我们首先使用二元关联(Binary Relevance)来实现一个简单的模型，原因为：其模型简单，易实现；可解释性强。主要目的是想将其作为一个baseline以便于同后续的模型进行对比。根据实验验证二元关联的测试结果并不理想，随后我们尝试分类器链模型和多标签逻辑回归模型(ML-LR)，其主要原因为：据分析，我们的标签之间相互是有关联的，而这两个模型都适用于二阶策略的多标签分类，能够在一定程度上考虑标签之间的关联性。

在深度学习算法上，我们首先通过微调BERT模型，其中预训练模型我们尝试了谷歌提供的原始BERT_Base_Chinese模型和哈工大讯飞联合实验室提供的BERT-Base-wwm [13]模型，使其适用于我们的任务。BERT-Base-wwm与BERT_Base_Chinese有两点区别，第一点是前者在训练过程中使用了基于全词Mask的训练样本生成策略。简单来说，原有BERT_Base_Chinese模型基于WordPiece的分词方式会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被mask。 在全词Mask中，如果一个完整的词的部分WordPiece子词被mask，则同属该词的其他部分也会被mask，即全词Mask [13]；第二点是BERT_Base_Chinese模型是以字为粒度进行切分的，没有考虑到传统NLP中的中文分词，而BERT-Base-wwm模型考虑了这点。随后，我们尝试引入BERT+RCNN模型，进一步提高系统的准确率，与微调BERT相同，我们使用了两个不同的预训练模型。

3. 类别不平衡

对于这个项目来说，对数据集进行分析可以发现，10000份样例20个分类标签中，约10个标签的样本数低于100，低于总样本数的百分1。通过这个实验可以发现这是一个极端的类别不平衡问题，通过过采样和降采样无法满足实验需求。因此在本项目中，我们小组打算退而求其次，只采用代价敏感学习技术以及决策输出补偿技术来针对这个数据集进行训练[14]。

系统架构设计

我们整个系统架构的设计如下图所示：

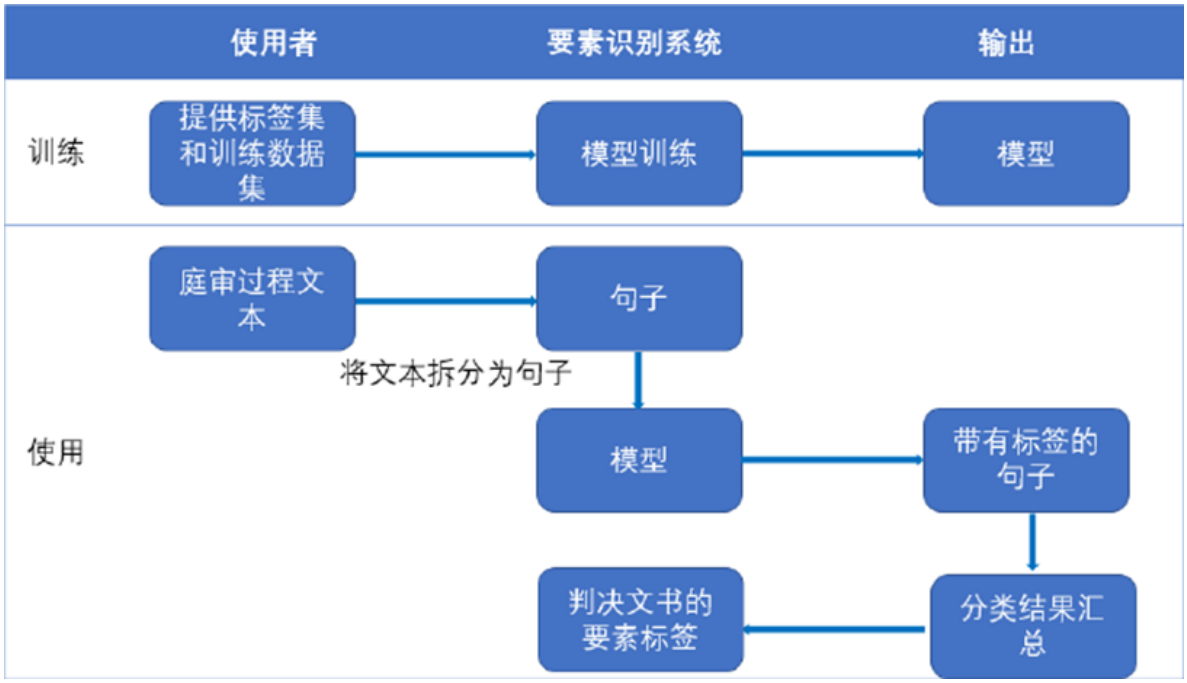

Figure.2 系统架构图

Figure.2中展示了本课题的两个部分的大致流程，分别是训练和预测。对于训练，在将输入的训练集投入练前，需要对数据进行预处理，如数据清洗，在训练完成后将模型保存;对于预测，通过载入对应的预测模型对输入的样本进行预测。

模型评估方法

由于我们选择的是2019年法研杯要素提取比赛的题目，所以我们选用的评估模型的方法与大赛官方一致，以便于和其他参赛者的结果进行比较。官方的模型评估方法如下：

- 对每种类型案件的模型分别使用对应的测试集得到对应的F1分数
- 对三个类型案件的F1分数取平均值作为最后结果

注：F1分数的计算方法如下：

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$
$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

同时，我们也使用准确率来评判模型，其公式如下所示：

$$Accuracy = \frac{\sum_{i \in TestSet} ||Predict\_Label_i == Real\_Label_i||}{\sum_{i \in TestSet} 1}$$
$$where\ ||Predict\_Label_i == Real\_Label_i|| = 1$$
$$if\ Predict\_Label_i == Real\_Label_i,$$
$$else\ ||Predict\_Label_i == Real\_Label_i|| = 0$$

目前结果

目前我们实现了Binary Relevance、Classifier Chain、ML-LR四种多标签分类机器学习算法，此外我们使用了BERT模型来做文本的多标签分类,我们分别对BERT进行微调和将BERT与RCNN结合来完成多标签分类的任务。由于在文献综述中对多标签机器学习算法做了说明，以下便不再赘述。

1. BERT 微调

BERT微调，是通过将我们的数据集输入到预先训练好的模型，从而改变预训练模型中的权重来适应我们的任务。我们采用的是Kaushal Trivedi对BERT集成的fast-bert包，它可以对二分类和多标签分类任务对预训练的语言模型进行微调。

2. BERT+RCNN

对于BERT+RCNN模型,它是将BERT的输出向量X输入到BiLstm,得到一个特征向量H,最后将X和H拼接送入Attention模块输出最后的预测结果。

以下是这几个模型的运行结果。模型最后的评判结果是模型在三种类型案件上预测的F1分数的平均值

1. 机器学习模型

	Binary Relevance	**Classifier Chain**	**ML-LR**
Divorce	0.00239118	0.05159578	0.44231245
Labor	0.0	0.03426911	0.38336548
Loan	0.0	0.03806323	0.37995813
Average	0.00079706	0.04130938	0.40187868

Table.1 机器学习模型F-1分数实验结果总表

2. 深度学习模型 (处理类别不平衡后)

	BERT(微调BERT_Base_Chinese)	**BERT(微调BERT-Base-wwm)**	**BERT+RCNN(BERT_Base_Chinese)**	**BERT+RCNN(BERT-Base-wwm)**
Divorce	0.4865353	0.6897	0.72173552	0.7927269
Labor	0.4425747	0.6162	0.61441835	0.7186667
Loan	0.4097913	0.6393	0.61687785	0.6974959
Average	0.4463004	0.6484	0.65101057	0.7362965

Table.2 深度学习模型F1分数实验结果总表

最后我们对两种在F-1分数上表现较好的模型进行了准确率的评估，其结果如下：

	BERT(微调BERT-Base-wwm)	**BERT+RCNN(BERT-Base-wwm)**
Divorce	0.7869	0.8280701
Labor	0.7531	0.8970432
Loan	0.8441	0.8028883
Average	0.7947	0.8426672

Table.3 两种深度学习模型准确率实验结果表

我们对BERT微调模型和BERT+RCNN模型均尝试使用事后决策输出补偿的方法来处理类别不平衡的问题，结果对比如下：

	BERT微调（决策补偿前）	BERT微调（决策补偿后）	BERT+RCNN（决策补偿前）	BERT+RCNN（决策补偿后）
Divorce	0.6455	0.6897	0.788703	0.7927269
Labor	0.5946	0.6162	0.696050	0.7186667
Loan	0.6012	0.6393	0.677152	0.6974959
Average	0.6138	0.6484	0.720635	0.7362965

Table.4 两种深度学习模型决策补偿前后F1分数

Table.4 中的模型使用的BERT模型均为BERT-base-wwm模型。

实验结果分析

从上述的实验结果可以看出,传统的多标签分类算法在我们这个问题上的表现很差,我们认为其主要的原因如下:

- 我们的数据存在严重的类别不平衡问题
- 我们采用的多标签算法多为first-order,即没有考虑标签之间的关系,导致最后在计算F1分数时会得到很低的分数。

BERT对谷歌提供的预训练模型没有哈工大科大讯飞联合实验室提供的预训练模型好，主要原因是在改变训练样本生成策略以及对中文进行词粒度上的切分后生成的预训练模型BERT-Base-wwm更能抓住中文语句的特征。

此外，我们对BERT微调模型以及BERT+RCNN模型均使用了决策输出补偿的技术来处理类别不平衡的问题，可以看出对模型效果的提升还不错。

附录:

人员分工

阮沛钧：BERT+RCNN 模型调试、训练

蓝文兴：微调BERT模型调试、训练

叶梓元：处理类别不平衡问题

时间安排

时间	内容
第一周	问题定义
第二周	多标签模型文献调研
第三周	句子向量化文献调研
第四周	二元关联算法实现
第五周	ML-KNN算法实现
第六周	数据集类别不平衡问题文献调研
第七周	第一次答辩
第八周	BERT微调模型
第九周	BERT+RCNN模型
第十周	模型F1分数对比及提高
第十一周	模型优化
第十二周	第二次答辩
第十三周	对BERT+RCNN进行类别不平衡处理
第十四周	处理BERT微调模型出现的问题
第十五周	优化模型效果
第十六周	第三次答辩

参考文献

[1] 胡发胜.要素模式引领审判权运行改革研究[J].山东人大工作,2015(11):21-24.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

[3] Rong Xin. word2vec Parameter Learning Explained.arXiv:1411.2738 [cs]. 2014.

[4] Le, Quoc, and Tomas Mikolov. Distributed representations of sentences and documents. International conference on machine learning. 2014.

[5] Lau, Jey Han, and Timothy Baldwin. "An empirical evaluation of doc2vec with practical insights into document embedding generation." arXiv preprint arXiv:1607.05368. 2016.

[6] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[7] Zhang M L, Zhou Z H. A review on multi-label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2013, 26(8): 1819-1837.

[8] Matthew R. Boutell,Jiebo Luo,Xipeng Shen,Christopher M. Brown. Learning multi-label scene classification[J]. Pattern Recognition,2004,37(9).

[9] Read J , Pfahringer B , Holmes G , et al. Classifier chains for multi-label classification[J]. Machine Learning, 2011, 85(3):333-359.

[10] Tsoumakas G , Katakis I , Vlahavas I . Random k-Labelsets for Multilabel Classification[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(7):1079-1089.

[11] ELKAN C. The foundations of cost0sensitive learning[C]. Proceedings of the 17th International Joint Conference of Artificial Intelligence, Seattle, Washington, USA, 2001:971-978.

[12] 于化龙. 类别不平衡学习理论与方法. 清华大学出版社. 2017.

[13] Y. Cui et al., "Pre-Training with Whole Word Masking for Chinese BERT," arXiv preprint arXiv:1906.08101, 2019.

[14] 周志华，杨强. 机器学习及其应用2011. 清华大学出版社. 2011.