非均衡数据下基于卷积神经网络的专利文本自动分类研究*

黄彩云[1] 吴金红[1] 陈勇跃[1] 王翠波[2]
（1. 武汉纺织大学管理学院，武汉 430200；2. 中南民族大学管理学院，武汉 430074）

摘 要：[**目的/意义**]探究非均衡专利文本数据的自动分类问题，并分析识别不同方案的分类效果。[**方法/过程**]使用卷积神经网络作为分类器，利用随机欠采样、随机过采样以及综合采样的方法对非均衡数据进行采样处理，使训练数据集均衡化，然后运用卷积神经网络进行自动分类，并分析分类的指标特征。[**结果/结论**]针对非均衡数据的分类问题，不能单一依据准确率来判定，三种实验方法都可以提高分类的正确率，但是进一步明确各类别实际的分类正确率而言，综合采样方法相对较好，能够改善专利文本自动分类效果。卷积神经网络在非均衡专利文本多分类中，虽然能够对多数类别进行相对较好的识别，但分类精度仍有较大提升空间。

关键词：卷积神经网络 非均衡数据 综合采样 专利自动分类
分类号：TP18
DOI：10.31193/SSAP.J.ISSN.2096-6695.2020.03.03

0 引 言

我国国家知识产权局的专利统计年报显示，从2016年至2018年，我国国内专利申请数量逐年递增[1]，这给我国相关专利审查部门的分析管理工作带来了巨大的挑战。面对海量的专利文本数据，有效的专利文本分类将会大大提高管理效率。目前国际上认可且应用比较广泛的专利分类依据为国际专利分类表（International Patent Classification，IPC）。IPC分类将与发明专利有关的全部技术内容按部、分部、大类、小类、主组、分组等逐级分类，组成完整的等级分类体系[2]。

* 本文系国家社会科学基金青年项目“基于实时大数据的潜在新兴技术敏捷预测机制研究”（项目编号：14CTQ017）研究成果之一。

[作者简介]黄彩云（ORCID:0000-0002-8249-1135），女，硕士研究生，研究方向为竞争情报、大数据分析，Email：1589222504@qq.com；吴金红（ORCID:0000-0003-2903-6372），男，博士，教授，研究方向为竞争情报、大数据治理及智能信息系统，Email：14514576@qq.com；陈勇跃（ORCID:0000-0002-5251-3516），男，博士，教授，研究方向为数据挖掘与信息可视化、信息与知识管理，Email：46258848@qq.com；王翠波（ORCID:0000-0003-1095-0897），女，博士，副教授，研究方向为技术竞争情报、数据挖掘与商务智能，Email：cbwang@mail.scuec.edu.cn。

通过专利分类，可以帮助相关的技术研发人员准确迅速地寻找相关的专利信息，提高二次创新的能力；帮助应用推广人员准确地将专利技术进行市场推广，有效地将技术落地并产生社会效益。但是，目前划分专利分类号的工作主要是依靠人工完成，而当前我国知识产权多个指标居于世界领先地位，专利数量的持续增长，给专利分类工作持续施压。因此，亟需引入先进智能的自动专利文本分类方法。

近年来，利用机器学习方法进行专利文本分类的研究也有不少，主要侧重点在于变换文本特征向量获得的途径后采用单一分类器进行文本自动分类，例如：Lim 等同时在标题、摘要、权利要求、技术领域和背景技术信息中抽取特征向量，进而改善专利文本分类效果[3]；Stutzki 等引入专利申请人地理位置特征，从而完善分类依据[4]；贾杉杉在专利自动分类研究的综述中表示，特征提取主要包括词袋模型、主题模型两种，分类器包括朴素贝叶斯、逻辑回归、神经网络、K- 近邻等[5]。随着深度学习方法的进一步发展，对专利文本分类器的研究也有新的突破，如：Zhu X 等在长短期记忆网络（Long Short-Term Memory，LSTM）的基础上构建了树状结构的模型 S-LSTM，利用 S-LSTM 中的记忆模块代替递归模型中的组合层，通过语义组合来理解文本[6]；Tai K S 等基于句法树的长短时记忆神经网络（Tree-LSTM），将标准 LSTM 的时序结构改为语法树结构，在文本分类上得到非常好的提升[7]；胡杰等提出了一种基于卷积神经网络与随机森林算法的专利文本分类模型，应用于英文机械专利文本分类场景[8]；Wang Y 等在 LSTM 的基础上引入 attention，以此捕获不同上下文信息对给定情感的重要性[9]；马建红等构建基于 attention 的双向 LSTM 模型（Bi-LSTM-ATT），对以 100 个专利应用效应作为类标签的机械物理类专利文本进行模型训练和分类测试[10]；Raffel C 等提出一种适用于前馈神经网络的简化注意力模型，证明了 attention 机制能够在文本较长的情况下，有效解决信息丢失等长期依赖问题[11]。

上述研究在训练集是均衡状态时均取得了较好的效果，但是如果数据集出现非均衡现象时，模型中的部分参数会发生偏移，其分类结果也会受到影响[12-15]。在实际的文本分类过程中，数据非均衡现象比比皆是，例如垃圾邮件、信用卡欺诈检测等等。针对专利文本分类的实际意义而言，考虑非均衡因素，能够有效避免专利的错误分类，对专利创新查漏补缺。因此，解决数据非均衡问题成为提高分类效果的另一研究热点。为提高专利文本分类的正确率，本文通过随机欠采样、随机过采样以及综合采样这三种方法对数据集进行处理后，选择卷积神经网络作为分类器对专利文本数据进行处理，通过对比实验，寻找专利文本自动分类下非均衡数据的最佳处理方法。

1　卷积神经网络与数据均衡

1.1　卷积神经网络

卷积神经网络是由 Hinton[16] 等人于 2006 年提出的一种深度学习模型。它是一种深度、多层前馈、反向传播的神经网络，目前在人脸识别、图像理解、计算机视觉、语言识别、无人驾驶等领域取得了出色的成就[17]。Kim 在运用卷积神经网络（Convolutional Neural Networks，CNN）处理文本分类过程中提出，模型主要包括输入层、卷积层、池化层及全连接层[18]。基本结构如图 1 所示。

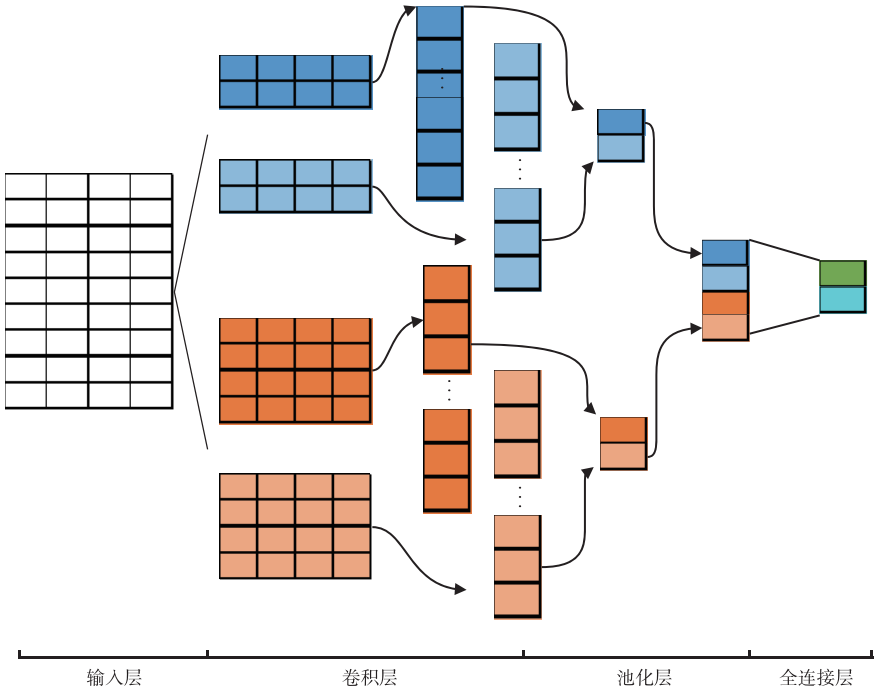


图 1　CNN 的基本结构

（1）输入层。主要完成数据的输入和预处理。在文本分类中，输入层将数据形成文本向量矩阵，每一个矩阵代表不同的句子。

（2）卷积层。用于提取文本的句子特征。其主要运行原理就是将卷积核与输入层的文本向量进行局部点积相乘，并按照一定的步幅从上至下依次以同一高度遍历输入层，从而获得输入层的特征映射。

（3）池化层。其主要作用就是全局缩小，也就是对卷积层形成的特征映射做进一步提取，将最重要的特征提取出来。由于卷积层所产生的卷积特征都是局部映射，存在大量不重要的样本，需要采用下采样的方法，进一步减少参数数量。

（4）全连接层。进一步对高度抽象化的特征进行整合，作归一化处理，针对上步得到的特征向量进行激活函数操作，得到每个文本所属不同类别的概率，为分类器进行判断提供依据。

1.2　非均衡数据下的 CNN 分类问题

在以往的研究中，CNN 能够在均衡数据中较好地对文本进行分类。然而，现实中的专利文本通常呈现出非均衡状态，主要表现在两个方面：一是类间的非均衡，具体表现为某一类或者多类样本数量明显少于其它类样本数量；二是在第一种情况下，单独某一个类别的样本中还存在类内不均衡现象，数据集内存在分离项目[19]。

从卷积神经网络的运行原理看，它是一种有监督的深度学习方法，将特征提取和类别判定集成在同一模型中统一实现，需要对数量接近的不同类别样本进行训练才能获得较好的泛化能力。当各类别中的样本数量差距较大时，模型会倾向于将类别判定为数据量大的样本，这样会使得损失率降低，从而不再需要继续优化参数。由此可见，CNN 对数据均衡性有一定要求，但在实

际中，所获得的分类数据往往不是均匀分布的，因此，非均衡状态的文本数据集利用 CNN 分类，其准确性会有很大偏差。

1.3　非均衡数据的处理方法

为了解决非均衡数据对 CNN 分类算法的影响，需要在进入 CNN 输入层之前对数据进行均衡化处理。通常有三类方法，包括随机欠采样（Random Under-Sample，简称 RUS）、随机过采样（Random Over-Sample，简称 ROS）以及综合采样（SMOTE+ENN）。

（1）随机欠采样法。对原始样本的多类样本中的数据进行部分抛弃，通过随机减少样本数量规模来达到整体样本数据集的均衡。

（2）随机过采样法。是和欠采样相对应的，对较少的对象进行补齐处理，通过复制少数类样本的策略来增加少数类样本的数量[20]。

（3）综合采样法。当实验中数据集的数量差距相对较大，单纯的欠采样或过采样都不能够解决正确率的问题，此时可以选取综合采样法，对两种极端情况的数据都进行相应的处理，从而使得整体数据达到相对均衡的状态。

2　非均衡数据的 CNN 专利文本分类模型

基于上述研究，我们构建了面向非均衡数据的 CNN 专利文本分类模型，如图 2 所示。

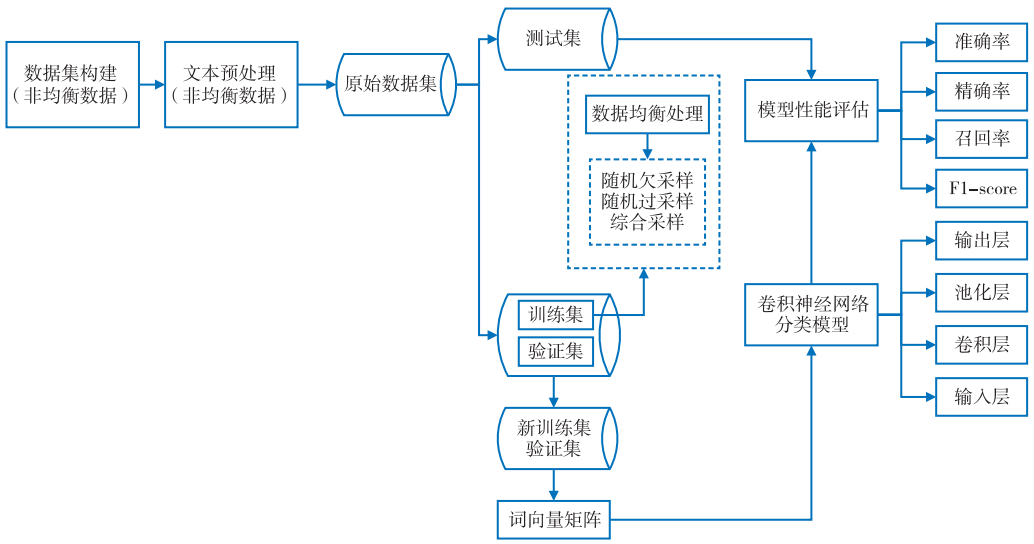


图 2　非均衡数据的 CNN 专利文本分类模型

2.1　数据均衡化

针对数据的欠采样处理，Phua 等提出了一种最近邻规则欠采样方法（Edited Nearest Neighbor，ENN）[21]，其主要思路是删除最近的 3 个近邻样本中的 2 个或者 2 个以上类别不同的样本。该方法的优点是可以使得整体数量变少，但是多数类样本附近的样本都是同类的，最终能

删除的样本数量十分有限，因而不一定能够提高分类正确率。

典型的随机过采样是采用合成少数类过采样技术（Synthetic Minority Oversampling Technique，SMOTE），算法的中心思想是随机线性插值的方法，从每个少数类样本的 K（K>1）种同类近邻样本中，挑选出最邻近的一个，在这两个样本间随机线性插值，得到新的人工少数类样本[19]。每一个随机选出的近邻 Pub_{ij} 分别与原样本 Pub_i 按照公式（1）构建新样本 Pub_{new}，$rand(0,1)$ 为 0 到 1 之间的随机数[22]。

$$Pub_{new} = Pub_i + rand(0,1) * (Pub_{ij} - Pub_i) \quad (1)$$

简而言之，SMOTE 方法是通过人工合成新的少数类样本，添加到样本中，使得各类样本数量达到均衡。然而，SMOTE 算法对每个原少数类样本产生相同数量的合成数据样本，而没有考虑其邻近样本的分布特点，使得类间发生重复的可能性加大[23]。

综合采样方法通常是将 SMOTE 算法和最近邻规则相结合，先利用 SMOTE 方法合成少数类从而扩充数据集，然后利用 ENN 方法清洗 SMOTE 过程中产生的噪点，合称 SMOTE-ENN 方法。当其中一种类型样本，与其最近的 3 个近邻样本中属于相异类型的样本数超过 2 个时，就删除这些相异类型样本，从而达到样本整体的均衡性。[22]

数据均衡化处理均用到 python 中的第三方库 imblearn。欠采样是利用 imblearn 库中的 RandomUnderSampler 函数，通过设置 RandomUnderSampler 中的 replacement=True 参数，实现自助法（boostrap）抽样。过采样利用 imblearn 库中的 SMOTE 函数，其中 kind 函数设置为 borderline1，表示最近邻中的随机样本与该少数类样本 a 来自于不同的类。调用 SMOTE-ENN 方法，可以直接实现综合采样。

2.2 文本向量的构建

一般来说，研究中收集的原始数据是文本的摘要部分，不能直接用作卷积神经网络的输入，必须经过数据的预处理。数据预处理过程包含两个部分：

（1）分词。本文使用 python 中的 jieba 来对文本进行处理，此种方法不仅可以根据专业要求自定义词典添加，同时对于没有添加且未登录的词，采用了汉字成词能力的隐马尔可夫模型（Hidden Markov Model，HMM），可以有效降低分词结果的不准确。

（2）去噪。剔除部分无用词组以及语气词等噪声词，以减少数据冗余，提高分词效果。

经过文本预处理，获得文档对应的分词结果，提取分词中的关键词。具体操作中，针对分词结果建立一定数量的 token 词典，将词典中的所有文字进行“数字列表”化，截长补短将所有“数字列表”转化为固定值，保证每个文本都是同样的长度。考虑到普通词袋模型会使得向量稀疏，这样卷积的大部分将是全零数字的情况，本文使用 keras 框架中自带的词向量层（Embedding），将“数字列表”转化为“向量列表”。文本向量矩阵表示为 $x=[x_1,x_2,x_3 \ldots x_n]$，其中 x_i 为第 i 个单词对应的词向量。

2.3 卷积层和池化层的算法

按照卷积神经网络的原理，将形成的词向量矩阵的结果作为卷积神经网络的输入，依次进行卷积、池化、再卷积、再池化循环，最后全连接至输出。各卷积层中的滤波器设置不同，但是同一层中的滤波器是权值共享的。特征映射的数量与卷积核的大小密切相关，计算方法为

$N=\frac{W-F+2P}{S}+1$。其中，W 为输入的文本大小；F 为感受视野的尺寸，即卷积核的大小；P 为移动到最后的边界补充值大小；S 为移动时候的步长。通过卷积层的作用，形成不同的特征信息，为池化层做准备。

池化层采用最大值池化。为防止过拟合，通常在全连接层之前加入 Dropout，以便随机丢弃隐含层的某些节点。使用基于梯度的优化器 Adam，采用 categorical_crossentropy 损失函数，同时生成准确度 accuracy 报告。

输出层中包含各类评价指标的值以及实验结果中具体各类别分类正确的混淆矩阵，根据结果能够具体判断模型性能的好坏，为进一步优化参数以及调整卷积神经网络的卷积层数提供依据。

2.4　模型评估方法

本文通过实验，来检验数据经过均衡化处理后，是否能够提高分类准确率。针对非均衡数据，在衡量各分类器性能时，传统方法中依靠准确率单一指标会掩盖对少数类别的不敏感性[15]，因此，本文采用混淆矩阵作为评价指标。常见的混淆矩阵结构如表 1 所示。

表 1　混淆矩阵

	预测真	预测假
真实真	真真（TP）	真假（FN）
真实假	假真（FP）	假假（TN）

由此矩阵构建如下几个评价指标，其中，准确率（Accuracy）指预测结果为正确的比率，精确率（Precision）用来检验是否查得准，召回率（Recall）用来检验是否查得全，F1–Score 值用来判断模型的稳健性。相关公式定义如下：

$$Accuracy=\frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$Precision=\frac{TP}{TP+FP} \quad (3)$$

$$Recall=\frac{TP}{TP+FN} \quad (4)$$

$$F1\text{-}Score=\frac{2TP}{2TP+FN+FP}=\frac{2*Precision*Recall}{Precision+Recall} \quad (5)$$

3　实证研究

3.1　数据源及数据预处理

本文选取专利 IPC 主分类中的部作为分类依据，各部的具体含义如表 2 所示。从 incoPat 全球专利数据库中，以纳米纤维为检索对象进行数据采集，采集内容包括专利标题、摘要以及主分

类号，共获得 115561 条带有标签的数据，经过筛选去重等处理后，剩余 114822 条数据，各类别数量见表 2。

表 2　数据来源及分类标签

IPC 部	技术含义	数量
A	人类生活必需	15862
B	作业；运输	19198
C	化学；冶金	27320
D	纺织；造纸	17592
E	固定建筑物	748
F	机械工程；照明；加热；武器；爆破	2004
G	物理	7530
H	电学	24567

本文的考察对象是专利数据库中专利信息的摘要部分，属于短文本分析。文本预处理后得到实验所需的原始数据集，为减少实验中出现过拟合现象以及因标签集中而影响分类效果，实验过程中利用 shuffle 函数将原始数据集自动打乱顺序。随后使用 train_test_split 函数将数据集自动划分为 80% 训练集、20% 测试集，其中只针对训练集进行数据均衡化处理。在词向量构建过程中，首先利用 jieba 对每条文本进行分词处理，利用 Tokenizer 模块对词进行编码，使每个词获得一个编号，将文本特征转换成数字特征。然后使用每个词的编号将每条文本转化为数字列表。由于每条文本的长度不一，依据以往经验，通过 pad_sequences 函数将每条文本长度固定为 50，超过部分截掉，不足部分用 0 补充。最后使用 Embedding 层将每个词编码转化成词向量。随后进行神经网络的各层级实验，词向量输入后进行多次卷积、池化后进行全连接，最后输出模型的准确率、精确率、召回率以及 F1–Score 值。

3.2　结果及分析

本文基于卷积神经网络，针对数据非均衡问题，分别进行了非均衡数据集的 CNN 文本自动分类实验和经过均衡化处理的数据集的 CNN 文本自动分类实验，所得结果如下。

3.2.1　非均衡数据集的 CNN 分类

将收集到的所有数据预处理后的非均衡数据集直接利用卷积神经网络进行实验，准确率只有 69.37%，实验过程中训练集和测试集的准确率变化以及经过 categorical_crossentropy 损失函数输出的损失值（loss）变化如图 3 所示。随着迭代次数的增加，训练集的准确率达到 95% 以上，但是测试集的准确率却不高，并且随着迭代次数的增加，训练集和测试集的 loss 值明显出现了过拟合现象，说明非均衡数据集的自动分类出现了问题。

查看混淆矩阵，对角线上的数值表示各类别的实际分类正确率，数据越大说明各类别的分类正确率越高，其结果如图 4 所示。可以看出，模型准确率不超过 70% 的情况下，各个类别的正

确率超过 70% 的有 A 和 H，但是类别 E 和 F 出现了严重的判断错误。查看数据源可知，E、F 类别是数据集中的小样本数据，考虑数据源层面可能是由于数据的非均衡性导致，使得数量少的类别被判错的概率变大。由于小样本数量过少，训练过程中特征提取不够，无法进行准确识别。同时，由于选择的数据源是同一领域，本身存在文本的相似性，可能存在 E、F 类别与其他类别的文本相似度较高，干扰分类效果。

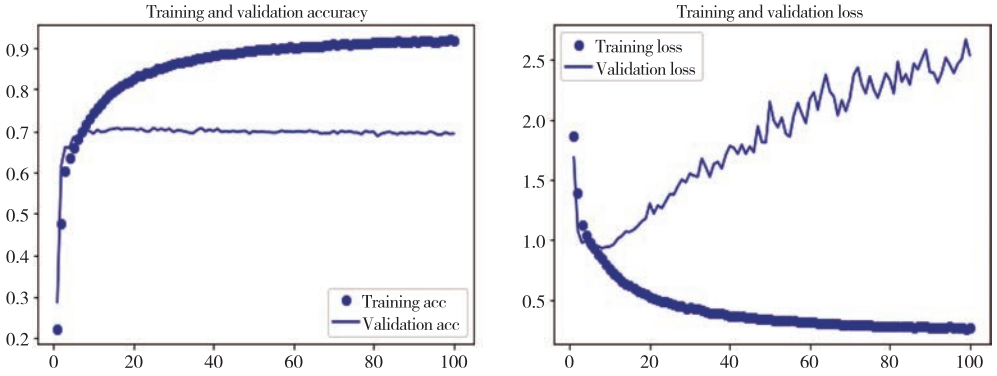


图 3　非均衡数据集中训练集和测试集中准确率和 loss 值变化图

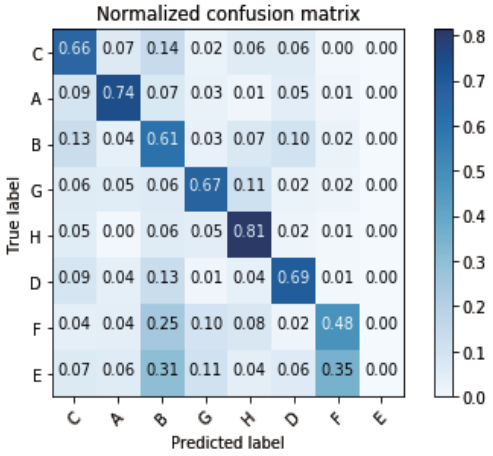


图 4　非均衡数据集的混淆矩阵

从非均衡数据集的实验结果可以判断出以下几个问题：（1）数据的均衡与否会影响实验的结果；（2）多分类问题中实验的评价指标不能依靠单一的准确率来衡量；（3）整体数据量较大的情况下，有数据类别不均时，数量过少的类别判断错误的可能性较大。

3.2.2　均衡化数据集的 CNN 分类

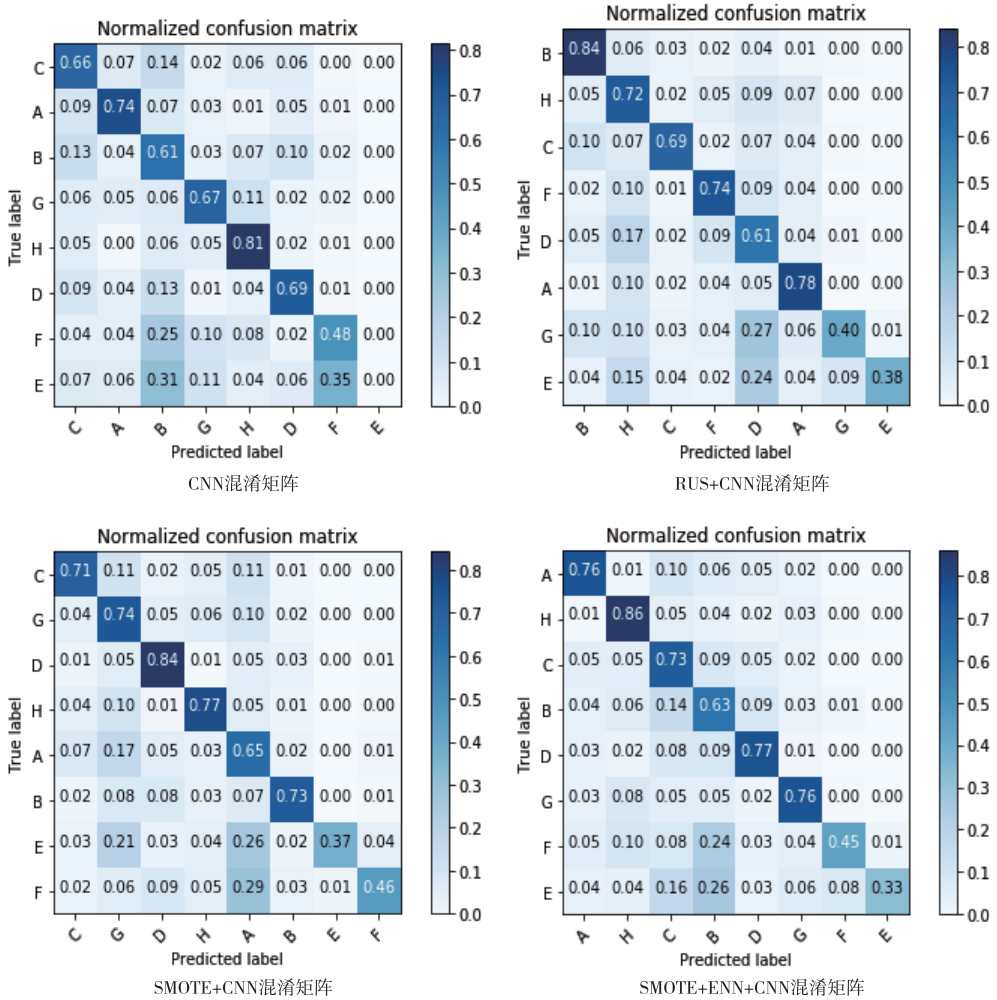
分别按照随机欠采样、随机过采样以及综合采样三种采样方法处理非均衡数据集，将处理后的实验结果与未处理的非均衡数据集的实验结果进行对比，如表 3 所示。

表 3　4 组实验结果对比

	准确率	精确率	召回率	F1–Score
CNN	0.6937	0.6940	0.6937	0.6932
RUS+CNN	0.7294	0.7284	0.7294	0.7282
SMOTE+CNN	0.7381	0.7406	0.7381	0.7383
SMOTE+ENN+CNN	0.7450	0.7105	0.6608	0.6779

未处理数据的非均衡问题会使得数量少的样本在自动分类过程中出现干扰，影响分类器的整体性能。相较于非均衡数据的实验结果而言，采用三种采样方法进行数据处理后，分类器的分类准确率、精确率均有所提升，其中综合采样的效果最好。由于数据源中大样本和小样本的规模相差近 30 倍，而且文本的相似度相较于不同来源的数据要高，所以存在整体实验 F1–Score 值不高，而且综合采样的召回率有所降低的情况。

同时查看 4 组实验的混淆矩阵，如图 5 所示。对比分析经过不同均衡处理的数据集中各个类别的预测正确率情况，很明显，经过均衡处理的实验结果中对角线颜色均变得较深。



CNN混淆矩阵　　RUS+CNN混淆矩阵

SMOTE+CNN混淆矩阵　　SMOTE+ENN+CNN混淆矩阵

图 5　4 组实验的混淆矩阵

根据混淆矩阵整理 4 组实验各类别的正确率，如表 4 所示。可以看出，虽然 3 种采样方法的整体正确率及其他三种评价指标都相对有所提升，但是从各类别的实际情况来看，RUS+CNN 主要是改善了少数类样本（E/F）的分类效果；SMOTE+CNN 在保证正确率提升、其他多数类别存在个别下降的情况下，整体性能提升不稳定；SMOTE+ENN+CNN 的准确率是最高的，基本保障各类别的预测正确率相较于未处理之前都有所提升。

表 4　4 组实验中各类别正确率汇总

	A	B	C	D	E	F	G	H
CNN	0.74	0.61	0.66	0.69	0.00	0.48	0.67	0.81
RUS+CNN	0.78	0.84	0.69	0.61	0.38	0.74	0.40	0.72
SMOTE+CNN	0.65	0.73	0.71	0.84	0.37	0.46	0.74	0.77
SMOTE+ENN+CNN	0.76	0.63	0.73	0.77	0.33	0.45	0.76	0.86

总体来说，综合采样方法弥补了欠采样过程中部分重要信息丢失和过采样中过拟合现象的不足，整体分类效果相对较好。

4　结　语

本文以专利文本的自动分类作为实证分析对象，针对使用传统的卷积神经网络技术对非均衡数据进行分类时存在的缺陷，分别采用随机欠采样、随机过采样以及综合采样三种方法对数据进行均衡化处理，然后利用卷积神经网络进行自动分类。将非均衡数据的分类实验结果和经过均衡化处理数据的分类实验结果进行对比，可以发现：针对非均衡数据的分类问题，不能单一依据准确率来判定；三种实验方法都可以提高分类正确率，但是进一步考察各类别的分类正确率以后发现，综合采样方法相对较好。需要说明的是，本文的文本分类处理，是经过文本分词后根据语义来判断，没有通过多个特征值去判定标签，其中存在一些噪声词影响分类效果；同时，专利文本中存在一定的专业名词，所以在词库的补充过程中也存在一定不足；另外，专利分类包含多个层级，本文只从部这一层级进行了多分类，没有进行更深层级的分类实验。这些均是本研究的不足之处以及需要加以改进的方向。

【参考文献】

[1] 国家知识产权局国内专利申请年度状况 [EB/OL]. [2019-08-02].http://www.cnipa.gov.cn/tjxx/jianbao/year2018/a/a3.html.

[2] 暴海龙，李金林 . 专利检索中的 IPC 和主题词识别方法研究 [J]. 北京理工大学学报 (社会科学版),2003,5(5):74-76.

[3] LIM S, KWON Y J. IPC Multi-label classification based on the field functionality of patent documents [C]//

International Conference on Advanced Data Mining and Applications. Springer, Cham, 2016: 677-691.

［4］STUTZKI J, SCHUBERT M. Geodata supported classification of patent applications［C］//Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data. ACM, 2016: 4.

［5］贾杉杉 , 刘小安 , 彭涛 . 基于 IPC 的专利文本自动分类研究综述［J］. 计算机科学 ,2017,44(10A):20-23.

［6］ZHU X, SOBIHANI P, GUO H. Long short-term memory over recursive structures［C］//International Conference on Machine Learning. 2015: 1604-1612.

［7］TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks［J］. arXiv preprint arXiv:1503.00075, 2015.

［8］胡杰 , 李少波 , 于丽娅 , 等 . 基于卷积神经网络与随机森林算法的专利文本分类模型［J］. 科学技术与工程 ,2018,18(06):268-272.

［9］WANG Y, HUANG M, ZHAO L. Attention-based LSTM for aspect-level sentiment classification［C］// Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606-615.

［10］马建红 , 王瑞杨 , 姚爽 , 等 . 基于深度学习的专利分类方法［J］. 计算机工程 ,2018,44(10):209-214.

［11］RAFFEL C, ELLIS D P W. Feed-forward networks with attention can solve some long-term memory problems［J］. arXiv preprint arXiv:1512.08756, 2015.

［12］TAN A C, GILBERT D, Deville Y. Multi-class protein fold classification using a new ensemble machine learning approach［J］. Genome Informatics, 2003, 14: 206-217.

［13］WEISS G M, PROVOST F. The effect of class distribution on classifier learning: an empirical study［J］. Technial Report ML-TR-44, Department of Computer Science, August 2, 2001.

［14］ESTABROOKS A, JO T, JAPKOWICZ N. A multiple resampling method for learning from imbalanced data sets［J］. Computational Intelligence, 2004, 20(1): 18-36.

［15］张文东 , 吕扇扇 , 张兴森 . 基于改进 BP 神经网络的不均衡数据分类算法［J］. 计算机系统应用 ,2017, 26(06):153-156.

［16］KRIZHEVSKY A,SUTSKEVER I, HINTON G. Image net classification with deep convolutional neural networks［J］. Communications of the ACM, 2017, 60(06):84-90.

［17］常亮 , 邓小明 , 周明全 , 等 . 图像理解中的卷积神经网络［J］. 自动化学报 ,2016,42(09):1300-1312.

［18］KIM Y. Convolutional neural networks for sentence classification［C］.//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.Doha, Qatar:EMNLP, 2014:1746-1751.

［19］陶新民 , 郝思媛 , 张冬雪 , 等 . 不均衡数据分类算法的综述［J］. 重庆邮电大学学报 (自然科学版),2013,25(01):101-110+121.

［20］CHAWLA N V, BOWYER K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique［J］. Journal of Artificial Intelligence Research,2002,16(1):321-357.

［21］PHUA C,ALAHAKOON D, LEE V. Minority report in fraud detection:Classification of Skewed data［J］. AcmSigkdd Explorations Newsletter,2004,6(1):50-59.

［22］李鑫 , 郭汉 , 张欣 , 等 . 基于非平衡数据处理方法的网络在线广告中点击欺诈检测的研究［J］. 计算机科学 ,2018,45(S1):371-374.

［23］WANG B X, JAPKOWICZ N. Imbalanced data set learning with synthetic samples［C］//Proc.IRIS Machine Learning Workshop.［s. l.］:［s. n.］, 2004.

文 献 与 数 据 学报
第 2 卷　第 3 期　　2020 年 9 月

Automatic Classification of Patent Text Based on Convolutional Neural Network under Unbalanced Data

HUANG Caiyun[1]　WU Jinhong[1]　CHEN Yongyue[1]　WANG Cuibo[2]

(1. School of Management, Wuhan Textile University, Wuhan 430200, China;
2. School of Management, South-Central Minzu University, Wuhan 430074, China)

Abstract: [**Purpose/significance**] This paper explores the automatic classification of unbalanced patent text data and analyzes the classification effects of different schemes. [**Method/process**] Using convolutional neural network as a classifier, using random under-sample, random over-sample and integrated sampling methods to sample unbalanced data, equalize the training data set, and then use a convolutional neural network for automatic classification and analyze the index characteristics of the classification. [**Result/conclusion**] For the classification of unbalanced data, it can't be determined solely based on the accuracy rate. All three experimental methods can improve the accuracy rate of classification , but to further clarify the actual classification accuracy rate of each category, the integrated sampling method is relatively good and can improve the effect of automatic classification of patent text. In the multi-classification of unbalanced patent texts, convolutional neural networks can recognize most categories relatively well, but there is still room for improvement in classification accuracy.

Keywords: Convolutional neural network; Unbalanced data; Integrated sampling; Automatic patent classification

（本文责编：王秀玲）