文章编号：1003-0077(2018)10-0069-09

基于多模型的新闻标题分类

董孝政，宋 睿，洪 宇，朱芬红，朱巧明

（苏州大学 江苏省计算机信息处理技术重点实验室，江苏 苏州 215006）

摘 要：该文研究中文新闻标题的领域分类方法（domain-oriented headline classification，DHC）。现有研究将 DHC 限定为一种短文本分类问题，并将传统分类模型和基于卷积神经网络的分类模型应用于这一问题的求解。然而，这类方法忽视了新闻标题的内在特点，即为“标题是建立在凝练全文且弱相关的词语之上的一种强迫性的语义表述”。目前，融合了序列化记忆的循环神经网络在语义理解方面取得了重要成果。借助这一特点，该文将长短时记忆网络模型（long-short term memory，LSTM）及其变型——门控循环单元（gated recurrent unit，GRU）也应用于标题的语义理解与领域分类，实验验证其性能可达 81％的 F1 值。此外，该文对目前前沿的神经网络分类模型进行综合分析，尝试寻找各类模型在 DHC 任务上共有的优势和劣势。通过对比“全类型多元分类”与“单类型二元分类”，发现在领域性特征较弱和领域歧义性较强的样本上，现有方法难以取得更为理想的结果（F1 值＜81％）。借助上述分析，该文旨在推动 DHC 研究在标题语言特性上投入更为充分的关注。

关键词：领域标题分类；卷积神经网络；循环神经网络

中图分类号：TP391　　**文献标识码**：A

Multi-model Based News Headline Classification

DONG Xiaozheng, SONG Rui, HONG Yu, ZHU Fenhong, ZHU Qiaoming

(Key Lab of Computer Information Processing Technology of Jiangsu Province,
Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: This paper studies the Domain-oriented Headline Classification (DHC) of Chinese news. The previous work defined DHC as a short text classification problem, and applied the classical classification model and convolutional neural network (CNN) model to solve the problem. However, these methods ignore the intrinsic features of the headline, i.e. a compulsive semantic expression based on the condensed text and weakly related terms. To exploit the adavantage of RNN in semantic representation, we apply the Long-short Term Memory (LSTM) and Gated Recurrent Unit (GRU) in DHC, achieving up to 81% F1-score. In addition, we systemically analyze the performance of the state-of-the-art neural network based classification models, with the purpose of revealing their common advantages and disadvantages for DHC. By comparing "multi-classification" to "binary-classification", we observed that the existing neural network models fail to achieve a performance better than 81% F1-score on the samples of strong domain ambiguity and weak domain characteristics.

Keywords: domain-oriented headline classification; convolutional neural network; recurrent neural network

0 引言

新闻标题领域分类（简称标题分类）的任务是依据标题语义对其所属领域进行判定，从而实现不同领域的标题归类。比如，标题“英国脱欧再现僵局”可归结为时政类领域。高效的标题分类对全文领域划分有着直接的促进作用，并因处理对象短而精，极大节省计算开销。因此，标题分类在基于领域特性的自然语言处理和计算语言学领域都有着重要的应用价值，比如垂直搜索、领域机器翻译[1]和舆情分析等。

收稿日期：2017-11-15　**定稿日期**：2018-04-02

基金项目：国家自然科学基金（61672367，61672368，61773276）；国防部科技战略先导计划（17-ZLXD-XX-02-06-04）

标题的主要特点是利用尽量精炼的语言概括丰富的信息。根据NLPCC2017共享任务中新闻标题分类的语料①的统计，95%的标题长度不超过20个汉字。因此，现有研究将标题分类归结为短文本分类问题。传统的短文本分类方法集中于分类规则和机器学习两个方面。

(1) 基于规则的方法源于专家知识的应用，依赖数据观测制定刚性的规则对标题类别进行界定，比如，如果标题中出现武器装备的术语或称谓，则判定其归属军事类领域。基于规则的方法往往处理速度快、精度高，但受限于观测数据的“小样本”现象，具有较低的泛化能力，在大规模数据处理时，往往暴露出较低的鲁棒性。

(2) 基于机器学习的短文本分类技术[2]则更为侧重分类模型的泛化能力，借助对标注样本的监督学习，优化分类核函数的目标并降低其误判损失，从而获得预定义特征集合上的判别模型。其泛化能力更强，可借助对训练样本类别特征的学习，对具有同类特征(同质异构)的测试样本进行判别。尽管如此，这类方法也继承了传统机器学习策略的不足，包括特征工程的手工化、学习过程的过拟合、训练数据的规模限制与类型分布失衡导致的偏差等。

目前，基于神经网络的深度学习技术已经获得重要突破，并在自然语言处理领域的诸多任务中取得良好成绩。这类技术在特征工程的去手工化、语义级特征学习与抽象、学习机制的抗数据及分布干扰方面，都具有潜在的优势。本文尝试将深度学习纳入短文本分类架构，应用于标题分类问题的求解过程，并结合标题成词造句的独特语言现象，验证各类技术的适应性和分析存在的缺陷。技术上，本文分别实现并应用了卷积神经网络(Convolutional Neural Network，CNN[3])、LSTM[4]和GRU[5]模型，以及结合了注意力机制的LSTM模型[6]，并借助这类模型对新闻标题中词向量(Embedding)形成多层语义感知。

实验中，基于CNN模型的标题分类评测，侧重检验不同局部语义特征及其联合的上层抽象对标题分类的贡献，这一实验的动机来源于如下发现：某些新闻标题中的词(或短语)呈现弱相关的特点，造句模式存在人为牵强的“拼凑”，比如，“新一代手机比纸薄能弯曲可卷折”。这类实例中，字、词或短语的独立含义对于类型判别有着更为直接的作用；相比而言，LSTM及其变形GRU更善于将词义及其相互关系进行建模，通过序列化的加权与融合形成统一的句子语义表示，与CNN模型在理论层面存在一定差异。比如，“茅房困境：买房不如买茅台”中，句子整体的语义有助于“茅房”(“茅台”与“房产”造作的合并)的词义理解，进而对全句的含义给予诠释，有助于这类实例的正确分类。因此，LSTM和GRU也作为本文重要的检验对象。此外，本文实验部分也检验了基于LSTM的注意力机制，意在评价词项的重要性对于全句分类的影响。

此外，本文建立了两套标题分类系统，一套为单模型多元分类系统，旨在利用一套深度学习模型解决多种领域的划分问题，另一套为多模型二元分类及投票系统，该系统联合使用CNN、LSTM和GRU三种模型对每一个标题样本进行单类型(是或非)的二元分类，对于不同模型在标题类型上具有不同判定的情况，将根据投票规则进行总体判断，包括“少数服从多数”和“确定性最高”两种原则。实验证明，简单的投票方式，可以将原有标题分类性能提高约1个百分点，且发现在实际测试中，某些标题样本无法被多种学习模型划分为任何一种类型，从而验证这类样本在多元分类系统强制的类型指派过程中，绝大部分难以避免误判。实验对这类标题样本进行了深入分析，并给出其主要特性。

总体，本文主要贡献如下：①检验了不同深度学习方法在标题分类中的性能；②提出并使用了多模型二元分类方法；③检验并分析了现有CNN、LSTM和GRU等模型对标题分类样本的适应性，并给出这类方法漏检的标题样本的特性。

本文组织如下，第1节简要介绍相关工作；第2节陈述标题分类的任务体系及数据资源；第3节介绍多模型二元分类系统架构，其中包括投票规则与实施方法；第4节还将给出CNN、LSTM及GRU的具体配置方法；第5节介绍实验架构及结果分析，其中包括漏检标题样本的特性分析，以及各类模型的适应性分析；第6节为总结全文及对未来工作的展望。

1 相关工作

传统短文本分类的方法主要涉及三方面工作：特征工程，特征选择和机器学习算法。在特征工程方面，最常用的特征是词袋模型(bag of word，BOW)，而其他复杂特征包括词性标签、名词短语、

① http://tcci.ccf.org.cn/conference/2017/taskdata.php

树核等。Post 等[7]将树核与不同任务的不同显示的树特征集合进行文本分类比较。特征选择即为删除“噪声”特征，提高分类的准确率，最常用的特征提取方法是移除文本中的停用词。相对而言，现有方法通过使用额外的知识进行知识扩展，弥补短文本特征少而又稀疏的缺点。Hu 等[8]通过利用内部和外部语义来提高短文本的性能。Banerjee 等[9]使用了维基百科数据扩充文本。机器学习算法则采用了逻辑回归(LR)、朴素贝叶斯(NB)和支持向量机(SVM)等分类模型。然而，这些方法均有特征稀疏的缺点，并且往往依赖于特定的场景和资源，难以进行推广。

最近，深度神经网络[10]和表示学习[11]在解决数据稀疏的问题上提供了新的思路，也提出了词表征的神经模型。Collobert 等[12]将卷积神经网络引入到了自然语言处理中的许多任务，并证明其提出的模型在各项任务中都获得了很好的表现。Kim[13]通过将单词向量与卷积神经网络进行结合，在短文本情感分类中取得良好效果。Santos 等[14]将英文短文本的字符序列作为处理单元，分别学习文本的词级和句子级特征，提高文本分类的准确性。这些工作证明了卷积神经网络在自然语言处理领域中有着广阔的应用前景，而循环神经网络通过使用带自反馈的神经元，能够处理任意长度的序列。因此，循环神经网络也已经被广泛地应用到自然语言处理任务中。

2　任务定义与评测体系

本文所涉研究继承了 NLPCC2017 系列共享任务中新闻标题分类(news headline category，NHC)任务的定义与评测方法。其中，标题定义为新闻网页中主体文字内容(非广告、推广或图像内容)的题目，分类标准为标题(仅仅考虑标题)语义呈现出的领域类型特性，比如，“茅房困境：买房不如买茅台”这一题目归为经济类。

根据任务定义，标题分类系统需对每一个标题样本给出唯一一个所属领域类别的标签，任务体系共给出 18 种领域类型(表 1)。按照这一分类体系，评测提供的人工标注样本数达到 228 000 条标题。其中，训练集包含 156 000 条标题，开发集和测试集各包含 36 000 条标题。此外，由于标注数据在不同类别上略显不均衡，因此，实际评测中，散文、故事、养生和探索四个类别的训练集、开发集和测试集分别各自包含 4 000、2 000 和 2 000 条标题；而除了这四个类别之外，其他的类别的训练集、开发集和测试集分别各自包含 10 000、2 000 和 2 000 条标题。

表 1　标题分类任务的领域类型体系

car(汽车)	baby(孩子)	game(游戏)	story(故事)
tech(科技)	world(世界)	sports (运动)	essay(散文)
food(食物)	military(军事)	finance(金融)	regimen (养生)
travel(旅游)	history(历史)	society(社会)	discovery(探索)
entertainment(娱乐)	fashion(时尚)		

针对任何分类系统给出的结果，评测统一采用精确率(precision)、召回率(recall)和 F1 测度进行性能评价。对 18 类整体的系统输出性能，则采用了微平均方法综合测评。

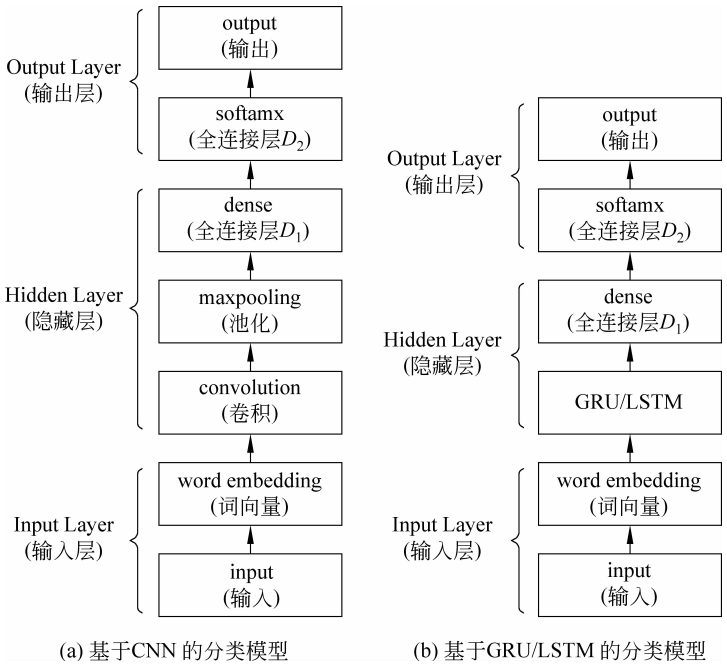
3　标题分类模型

本节介绍基于深度学习的标题分类，包含两个部分：分类模型结构和多模二元分类投票机制。

3.1　分类模型结构

本文采用多层深度感知的神经网络构建标题分类模型。其基本架构如图 1 所示，包括输入层、输出层和隐藏层。基于 CNN、LSTM 和 GRU 的分类模型的输入层、输出层具有相同结构。在输入层都接收来自某一标题中所有词的词向量，而在输出端则给出不同类别上的概率分布。显然，实践过程中，最大概率的类别标记将作为输入标题样本的分类结果进行输出。

上述三种分类模型的区别主要集中在隐藏层，基于 CNN 的分类模型，在隐藏层增设了卷积层和最大池化层，而 LSTM 和 GRU 则在隐藏层嵌入了递归神经网络，并增设记忆控制门等门控机制。上述分类模型的性能优劣，主要取决于隐藏层内的学习方式：独立特征优先(CNN)或序列语义优先(LSTM 或 GRU)。



(a) 基于CNN 的分类模型 (b) 基于GRU/LSTM 的分类模型

图 1 标题分类模型架构

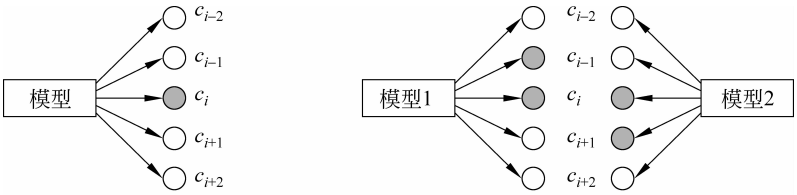
GRU 本身是 LSTM 的一种简化变形，但在使用过程中，对于序列语义的选择性记忆略显优势，因此，本文也将其作为一个独立的方法，参与上述分类的对比评价，以及三者的投票机制。关于隐藏层中的 dense 和 Dropout 两层的配置，根据不同的学习模型而略有不同，但作用并无差异。

3.2 多模二元分类投票机制

针对特定标题样本，借助神经网络的多层感知，分类模型可在网络的输出层对所有可选领域类别进行概率指派，比如，本文以 NLPCC2017 共享任务体系中分类任务的 18 种领域类别进行判别，则分类模型在输出层利用 softmax 回归模型计算每种类别的概率，最高概率对应的类别标签将作为输入样本所属的领域的解。但是，在实际执行过程中，某些样本在领域划分上存在一定的模糊性，例如，“马云演讲视频《CCTV 创业英雄会》”既可以认为是一种科技领域的题材（依据马云背景），也可认为是娱乐领域的题目（依据 CCTV 节目秀）。那么，建立一个单模型多元（18 元）分类的系统，进行刚性的唯一类别指派，往往形成过于武断的判别机制。

为此，本文尝试结合多个分类模型的二元分类结果（是或非），利用规则形成简单的投票方法，形成较为灵活的判别机制。图 2 中的左子图（a）是传统的单模型多元分类机制，其中，在单一神经网络的输出层，经过 softmax 回归后，18 种领域类别中，只有概率最高的一种类别作为结果输出；相比而言，右子图（b）则建立了两套分类模型，每套模型对单一标题样本的 18 种可能的领域类别全部进行二元判别（是为 1，否则为 0），从而每套模型都可能将某一样本划分到多种领域类别之中。那么，两套分类模型对这一样本的二元类型判别结果，就形成了两套不同的领域类别集合，两者或许存在交集（图例中给出的是存在交集的情况），或许互不统一。而前文提到的规则，即是在上述不同模型得出的领域类别集合之上实施的进一步判断。在本文实际的实验中，我们引入三种深度学习模型（CNN、LSTM 和 GRU）进行二元判定，由此，规则实际上是在三者得出的三套类别集合上实施的综合判断。

针对多模型二元分类产生的类别指派，本文采用“少数服从多数”和“可信度最高”两项规则进行二次筛选，将较为可靠的类型指派给目标标题样本。注意，之所以未保留多于一项的类型指派，原因在于 NLPCC2017 分享任务中标题分类语料的单一类别标记原则，即每个标题只由人工指派唯一一个类型标签。那么，上述多模型二元分类产生的多种类别产出的现象，可认为不同模型根据自身不同的理论基础得出的不同中间结果，而结合不同模型得出的判断，再基于规则得出综合判断的过程，可理解为一种对“多候选择优的人工单选题求解方式”的模拟。比如，基于独立词义特征进行深度学习的二元 CNN 模型，可对领域信息混杂的标题样本实现多个无关



(a) 单模型多元分类
(判断类别仅为c_i)

(b) 多模型二元分类
(判断类别可能为c_{i-1}、c_i和c_{i+1})

图 2　单模多元分类机制与多模型二元分类机制对比

(或弱相关)领域类型的临时判断,假设标题“什么才是真正自动驾驶？开车打一局王者荣耀才算”中的“自动”“驾驶”和“王者荣耀”分别利于科技、汽车和游戏领域类型的识别,且 CNN 确实输出上述类别标签;同时,综合了全局语义的 LSTM 模型,对上述标题也给出了汽车和散文两种模糊的类型判断。那么,本文规则的目标即是综合评定两者的判断,给出汽车领域这一最终判断(少数服从多数原则)。

本文具体的规则如下所示:

规则 1　三套二元模型中至少有一套模型有输出结果(图 3 展示了每套模型含有多个结果的情况),根据输出结果统计 18 个领域出现的次数$[\mathrm{num}_1,\mathrm{num}_2,\cdots,\mathrm{num}_{18}]$(18 个领域分别为 c_1,c_2,…,c_{18}),将对应的出现次数由大到小排序$[\mathrm{num}_m,\mathrm{num}_n,\cdots]$,其中 m、n 分别对应原始次数中的下标。标题领域 label 确定如式(1)所示。

$$\mathrm{label}=\begin{cases}c_m,\mathrm{num}_m>\mathrm{num}_n\\ \text{无标签，否则}\end{cases}\tag{1}$$

规则 2　对于规则 1 中无标签的结果,则采用高置信度的模型(GRU)的多元分类模型重新分类,且唯一输出多元分类的单一判定结果。

规则 3　当所有模型都没有输出结果时(即每个二元分类模型在 18 种领域类型上都判定为 0,即无关),则最终认定标题样本不属于任何领域类型。

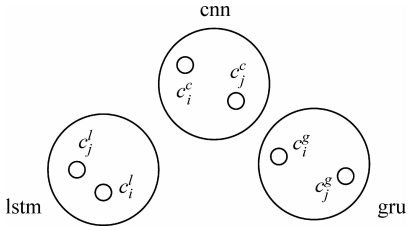


图 3　多模型二元分类多输出

上述规则中,最后一项规则是不符合 NLPCC2017 分享任务中标题分类评估要求的输出方式,因为任务要求每一个样本都唯一对应一个领域类型,不为空。因此,上述最后一项空输出的规则,必然导致召回率 R 的下降,实验中也的确显示了这一现象(尽管最终系统性能因为精度 P 的提升获得了约 1%$F1$ 值的提高)。我们将这一规则纳入考虑的动机在于,经过人工观测,某些标准数据集中的标题样本并不属于指定的 18 项领域类别,人工标记存在牵强的指派。为此,本文利用空输出的规则,收集疑似的错误标记样本并进行分析,如实验分析部分的介绍(第 5 节),部分空输出的样本的确存在误标记,但另外一部分样本则暴露了现有分类模型的不足。

4　深度学习模型的配置

本节介绍神经网络模型的参数配置,包括基于 CNN、LSTM 和 GRU 的多元分类模型和二元分类模型。对于单模多元分类模型,我们只需构建一个多(18)元分类器,对于二元分类模型,我们需要对每种领域构建对应的二元分类器(参数相同),即二元分类模型包括 18 个二元分类器。

4.1　CNN

对于 CNN 的多元分类器和二元分类器,输入一个新闻标题(已通过 jieba 工具分词),通过向量层将新闻标题转化为向量。然后通过卷积和池化提取标题特征,卷积和最大池化伪代码如下:

输入:新闻标题向量 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$
输出:新闻标题特征向量

```
repeat
  for 每个卷积核 ω do
    for i in n−h+1 do  //卷积核的宽度为 h
    o_i ← tanh(ω · x_{i:i+h−1} + b_o)
    end for
    c ← concat(c, o_i)
  end for
  c′ ← maxpooling(c)
until 输出 c′
```

基于CNN的多元分类器的参数配置如下：①卷积核窗口大小为3，卷积核的数量为128；②全连接层D_1单元数为64，激活函数为tanh，全连接层D_2单元数为18，激活函数为softmax；③Dropout为0.5。而基于CNN的二元分类器参数配置如下：①卷积核、全连接层D_1和Dropout的设定与基于CNN的多元分类器相同；②全连接层D_2单元数为1，激活函数为sigmoid。

如表2所示，CNN模型训练参数，二元分类器的损失函数是Loss_1，而多元分类器的损失函数是Loss_2[①]，二元分类器和多元分类器的优化器和迭代次数设置如表2所示，且设置的最小批数据是256。

表2 CNN模型训练参数

Activation Function（激活函数）	Loss Function（损失函数）
$\text{sigmoid}(x)=\frac{1}{1+e^{-x}}$	$\text{Loss}_1=-\log P(Y\mid X)$
$\text{softmax}(a)_j=\frac{e^{a_j}}{\sum_{k=1}^{K}e^{a_k}}$	$\text{Loss}_2=-\sum_j y_j\log p_j$
$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	
Optimizer(优化器)：Adam	Iterations（迭代次数）：20次

4.2 LSTM和GRU

基于LSTM的模型和基于GRU的模型，其输入与基于CNN的模型相同，二者都是时序化处理标题得到标题语义向量。LSTM隐层的伪代码如下：

输入：数据$\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$，$x_t$是标题第$t$个词向量
输出：时序化处理标题，得出标题语义向量

```
repeat
    对于当前步，输入 x_t、前一层的隐含层输出 h_{t-1} 和细胞状态 C_{t-1}
    //忘记门层 f_t
    f_t ← sigmoid(W_f · [h_{t-1}, x_t] + b_f)
    //输入门层 i_t
    i_t ← sigmoid(W_i · [h_{t-1}, x_t] + b_i)
    //新候选值向量 C̃_t
    C̃_t ← tanh(W_c · [h_{t-1}, x_t] + b_c)
    //信息融合新的细胞状态
    C_t ← f_t * C_{t-1} + i_t * C̃_t
    //一个确定细胞状态的输出的值
    o_t ← sigmoid(W_o · [h_{t-1}, x_t] + b_o)
    //隐含层输出 h_t
    h_t ← o_t * tanh(C_t)
until 遍历所有的词，输出 h
```

GRU是LSTM的变种，GRU隐层的伪代码如下：

输入：数据$\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$，$x_t$是标题第$t$个词向量
输出：时序化处理标题，得出标题语义向量

```
repeat
    对于当前步 x_t，会存在前一层的隐含层输入 h_{t-1}
    z_t ← sigmoid(w_z · [h_{t-1}, x_t])    //更新门
    r_t ← sigmoid(w_r · [h_{t-1}, x_t])    //重置门
    h̃_t ← tanh(W · [r_t * h_{t-1}, x_t])   //候选输出
    h_t ← (1 − z_t) * h_{t-1} + z_t * h̃_t  //当前步的输出
until 遍历所有的词，输出 h
```

通过GRU和LSTM处理文本得到标题语义，输出维度为128维。另外，基于GRU的多元分类器和基于LSTM的多元分类器的其他参数配置与基于CNN的多元分类器相同。

5 实验

5.1 词向量

我们使用了预训练的Word2vec模型[②]，词向量[15]维度为256维，由Gensim训练得到。在标题分类任务中，若单词不存在预训练的向量词集中，则进行随机初始化。此外，标题预定长度为40，若长度大于预定长度则舍去超过部分，若小于预定长度则进行补零操作。

5.2 实验设置

单模多元分类模型和二元分类模型（第4节，实验配置） 对于每套模型的二元分类模型的预测结果，只有单输出结果才能确定标签。不同网络的二元分类模型的预测结果，在测试集中能确定标签的数据不同，定义CNN二元分类模型能够确定标签的部分测试集数据为Data1，LSTM和GRU二元分类模型能确定唯一标签的数据集分别是Data2和Data3，整个测试集数据为Data。

LSTM+Attention 对于LSTM的多元分类模型，我们在时序化处理文本向量时添加了注意力机制，用于新闻标题分类。

融合系统 通过投票策略（详见3.2节）融合CNN、GRU和LSTM三种二元分类模型，即多模二元模型。因为投票策略中第三条规则的数据不符合

① keras-cn.readthedocs.io/en/latest/

② http://spaces.ac.cn/archives/4304/comment-page-1

评测任务要求，强行使用 GRU 的多元分类对没有输出结果的数据也强行指派了一个分类。

融合系统前提是 Data1、Data2 和 Data3 数据集不能完全一致，三种数据集的重叠部分如图 4 所示。每个序号对应的数量如表 3 所示，其中 5、6 和 7 分别对应 Data1、Data2 和 Data3。从数据结果来看，三个数据集具有差异，从而保证了融合系统的可实施性，也验证了不同模型对标题分类任务的适应性是有区别的。

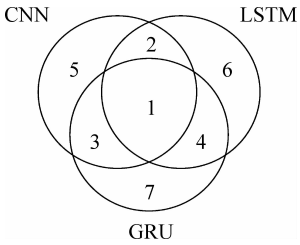


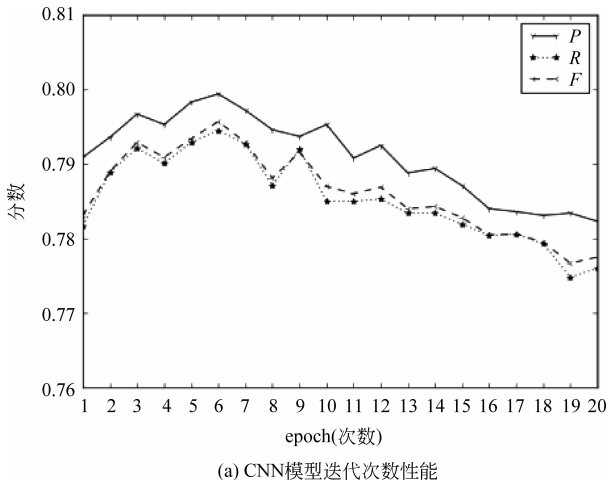
图 4　数据集重合图

表 3　序号相应数据数量

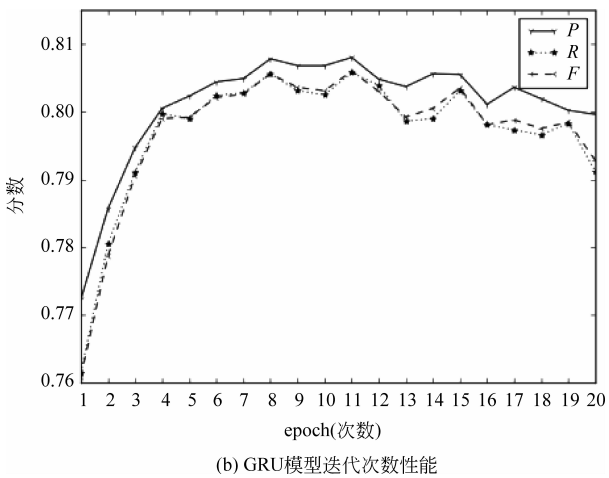
序号	数据量	序号	数据量
1	22 248	5	28 169
2	24 165	6	28 381
3	24 542	7	28 775
4	25 029	—	—

5.3　模型的最优迭代次数选

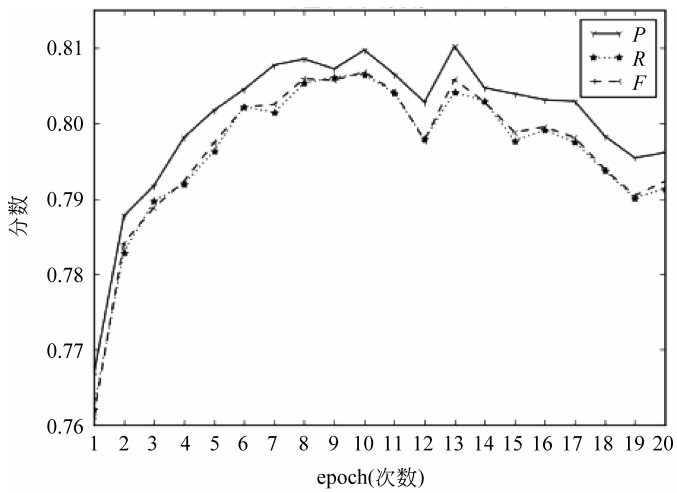
我们使用了 36 000 条开发集数据，根据图 5，各模型的 P、R 和 F_1 值与迭代次数的关系图所示，来选择对应的性能最优的单模多元分类模型。因此，我们分别得到了基于 CNN、GRU 和 LSTM 的多元分类模型的对应最优迭代次数为 6、8 和 10。因为二元分类模型存在部分数据无法指定其标签，所以选择迭代模型时参照了单模多元分类的迭代次数的性能。



(a) CNN模型迭代次数性能



(b) GRU模型迭代次数性能



(c) LSTM模型迭代次数性能

图 5　P、R 和 F_1 值与迭代次数的关系图

5.4 结果与分析

多元分类模型和二元分类模型在对应数据集上的分类结果对比如表4所示。

表4 多元分类与二元分类模型对比

	模型	P/%	R/%	F_1/%	数据集
多元分类模型	CNN	86.15	85.93	85.97	Data1
	LSTM	87.06	86.93	86.93	Data2
	GRU	86.94	86.83	86.83	Data3
二元分类模型	CNN	86.97	86.78	86.68	Data1
	LSTM	87.33	87.15	87.15	Data2
	GRU	87.69	87.54	87.56	Data3

根据表4结果，在相应数据集上，CNN、LSTM和GRU的二元分类模型的微观精确率、召回率和F_1值都高于多元分类模型，因此二元分类模型性能优于多元分类是融合系统性能提升的因素。

同时，我们观测了三种网络的多元分类模型预测结果的错误样例，发现三种网络对标题分类任务的适应性是有区别的，单模多元分类预测错误样例如表5所示。

表5 单模多元分类预测错误样例

单模多分类模型	实例	真实标签	预测标签
CNN	*世界上最令人不解的5大未解之谜*	探索	世界
	五种人不适宜喝鸡汤	养生	食物
LSTM	*马路上的“爱情”*	社会	文章
	用好一线一指标	金融	军事
GRU	*新能源汽车龙头的三个选择*	金融	汽车
	DNA检测不是产前诊断“金标准”	孩子	科技

对表5中多元分类模型预测的错误样例依次进行分析，在前两个样例中，根据“世界”特征预测为“世界”标签，根据“鸡汤”误分类为“食物”标签，仅仅根据局部特征进行分类，而忽略了标题的语义。虽然LSTM可以很好地利用时序化信息，但由于“马路上的爱情”和“一线一指标”这两个样例缺少上下文信息，导致LSTM依然不能将其正确分类。最后两个样例中，“汽车”和“DNA检测”的特征在整个标题语义中占据比重较大，导致分类错误。根据上述样例分析，当标题存在明显特征时，基于CNN的多元分类模型更有优势；而基于GRU和LSTM的多元分类模型通过时序化处理标题，得到整个标题语义的表示，进行标题分类，更适应需要理解标题语义的数据，三个模型对标题适应性的不同是二元模型确定标签数据集合不同的内在体现。

融合系统与其他实验对比如表6所示，其中也增添了NLPCC评测任务中排名靠前的结果（因为论文未公开，只使用结果），同时使用了整个测试集Data。

表6 实验结果

模型		P/%	R/%	F_1/%
多元分类模型	CNN	79.61	79.04	79.18
	LSTM	80.48	80.15	80.18
	GRU	80.63	80.28	80.32
NLPCC	Top	83.11	82.97	83.04
	Second	82.82	82.56	82.69
	Third	81.86	81.49	81.68
	Fourth	81.65	80.96	81.30
LSTM+Attention		78.60	78.48	78.42
融合系统		**81.41**	**81.14**	**81.16**

根据表6结果，我们发现添加了注意力机制的LSTM多元分类效果不佳，没有单独的LSTM多元分类模型好，主要因为标题是凝练全文语义的简要描述，而该模型重点关注局部词的特征，会改变整个标题语义的表示。而使用投票规则的融合系统，P、R和F_1值都要高于较优的基于GRU的多元分类模型，且召回率提升了约1%的性能，证明了我们的融合系统方案在新闻标题分类任务上是有效的。NLPCC评测任务以召回率为评价标准（系统没有公开，只对比了结果），与第一名相比，R值低了1.5%，与第三名相当，虽然没有达到最优，但是说明了系统的性能是可靠的。此外，通过融合系统能够找到一些漏检数据（通过GRU多元分类模型指定分类的数据）。

如下面的漏检例子所示，括号中的标签为漏检数据样例的真实标签：

① 印度骆驼节狂欢场面壮观（世界）；

② 鸠兹古镇居然是个吃货镇（食物）；

③ 应知应会三字经免费拿（养生）；

④ 用洪荒之力查出泄密黑手（金融）。

前两个样例的预测结果有多个标签，后两个样例无预测标签。多模二元分类模型将第一个示例预测结果为“世界”和“旅游”标签，而第二个示例被预测为“食物”和“旅游”标签，根据两个实例的语义和

词义，可以将其划分到“旅游”领域，因此将类似的数据归为强语义歧义特征数据。在后两个示例中，由于标题中的特征较少，因此分类模型将其分到正确领域是较为困难的，多模二元分类模型仅依靠标题中的语义不能将其归属于任务中一个领域，将这些数据归为弱语义特征数据。通过表 5 和表 6 对比，多元分类模型性能的下降验证了强语义歧义特征和弱语义特征的数据对标题分类影响很大。

6　总结和展望

本文尝试了将神经网络的二元分类用于多分类目标预测，并且选用了多种结构的神经网络模型。利用投票机制将 GRU、CNN 和 LSTM 的二元分类模型形成的融合系统用于标题分类任务，评测分数中召回率达到 81.14%，比最优的 GRU 多元分类提高了约 1%的性能。另外，本系统能够区分强语义特征和弱语义特征的数据，对于分类具有重要意义。

未来工作安排主要包括两个方面：①判断强语义歧义特征数据的类别归属；②对弱语义特征数据添加额外信息增加弱语义特征数据的可区分性，比如添加正文中的高频词、实体和词性等信息。

参考文献

[1] 刘昊，洪宇，姚亮，等. 基于 HITS 算法的双语句对挖掘优化方法[J]. 中文信息学报，2017，31(2)：25-35.

[2] 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报，2004，18(1)：26-32.

[3] Krizhevsky A，Sutskever I，Hinton G E. Imagenet classification with deep convolutional neural networks [C]//Proceedings of Advances in Neural Information Processing Systems. 2012：1097-1105.

[4] Hochreiter S，Schmidhuber J. Long short-term memory[J]. Neural computation，1997，9(8)：1735-1780.

[5] Chung J，Gulcehre C，Cho K H，et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv：1412.3555，2014.

[6] Rocktäschel T，Grefenstette E，Hermann K M，et al. Reasoning about entailment with neural attention[J]. arXiv preprint arXiv：1509.06664，2015.

[7] Post M，Bergsma S. Explicit and implicit syntactic features for text classification[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2：Short Papers)，2013，2：866-872.

[8] Hu X，Sun N，Zhang C，et al. Exploiting internal and external semantics for the clustering of short texts using world knowledge[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM，2009：919-928.

[9] Banerjee S，Ramanathan K，Gupta A. Clustering short texts using wikipedia[C]//Proceedings of the 30th Annual Inter national ACM SIGIR Conference on Research and Development in Information Retrieval. ACM，2007：787-788.

[10] Hinton G E，Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science，2006，313(5786)：504-507.

[11] Bengio Y，Courville A，Vincent P. Representation learning：A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(8)：1798-1828.

[12] Collobert R，Weston J，Bottou L，et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research，2011，12(Aug)：2493-2537.

[13] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv：1408.5882，2014.

[14] dos Santos C，Gatti M. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proceedings of COLING 2014，the 25th International Conference on Computational Linguistics：Technical Papers，2014：69-78.

[15] Mikolov T，Sutskever I，Chen K，et al. Distributed representations of words and phrases and their compositionality [C]//Proceddings of Advances in Neural Information Processing Systems. 2013：3111-3119.

董孝政(1994－)，硕士研究生，主要研究领域为自然语言处理、文本分类。

E-mail：xiaozhengdong@yahoo.com

洪宇(1978－)，通信作者，副教授，硕士生导师，主要研究领域为话题检测、信息检索和信息抽取。

E-mail：tianxianer@gmail.com

宋睿(1993－)，硕士研究生，主要研究领域为自然语言处理、文本分类。

E-mail：cnsr27@gmail.com