文章编号：2095-2163(2020)08-0242-07 中图分类号：TP391 文献标志码：A

基于 LSTM 的《红楼梦》文本风格分界点识别方法

朱东旭，严广乐
（上海理工大学 系统科学系，上海 200093）

摘 要：针对《红楼梦》文本风格分析中，手工设计的特征带有主观性与仅根据单个分界点前后的差异直接判断分界点位置的问题，提出了一种基于深度学习的文本风格分界点识别方法及用于文本分类的深度模型 1-DconvLSTM。该方法在不同的训练分界点处，相互独立的训练多个结构相同的 1-DconvLSTM 模型。通过测试模型的泛化性能，得出在不同训练分界点处前后差异显著性的大小，并据此确定文本风格的实际分界点位置。实验结果表明，红楼梦的文本风格分界点在第 80 回，强有力的支持了《红楼梦》80 回前后不是同一作者所著的结论。

关键词：红楼梦作者分析；文本风格分析；词向量；自然语言处理；1D-CNN LSTM

A method of identifying the style demarcation point of A Dream of Red Mansions based on deep learning

ZHU Dongxu，YAN Guangle
（department of systems science，University of Shanghai for Science and Technology，Shanghai 200093，China）

【Abstract】 Aiming at the problems of manual design features in traditional A Dream of the Red Mansions text style analysis with subjective issues and directly judging the position of the style demarcation point based on the difference between the single demarcation point before and after，this paper proposes a text demarcation point recognition method based on deep learning and a text classification model 1-DconvLSTM. First，the method trains multiple 1-DconvLSTM models with the same structure independently at different training cut-off points，and then tests the generalization performance of these models to obtain the magnitude of the significant difference between the different training cut-off points. Based on this，the actual demarcation position of the text style is determined. The results of the experiments show that the text style demarcation point of A Dream of the Red Mansions is at 80 chapters，which strongly supports the conclusion that A Dream of the Red Mansions is not written by the same author around 80 chapters.

【Key words】 author analysis of A Dream of the Red Mansions；text style analysis；natural language processing；1D-CNN；LSTM

0 引 言

《红楼梦》（也称《石头记》，以下简称“红楼梦”）作为中国四大名著之首，无疑是历史上最具文学价值的作品之一。围绕红楼梦进行的各种研究，尤其是关于红楼梦作者的研究，一直受到广大学者的关注，各种说法层出不穷。

早期的红楼梦作者分析中，主要是针对手工设计的特征进行统计分析。1981 年陈炳藻教授采用统计学方法，利用 t 检验，得出红楼梦 120 回全是曹雪芹一人所作的结论[1]；1987 年李贤平教授研究了 47 个虚词在各章节的分布规律，并使用聚类方法，得出前 80 回与后 40 回不是同一作者的结论[2]；2009 年韦博成教授提出以花卉、树木、饮食、医药和诗词 5 个场景的出现频率为指标，运用统计学中“两总体等价性检验”的理论和方法，得出红楼梦前 80 回与后 40 回在某些重要情节的描写上确实存在显著差异的结论[4]。

近期的红楼梦作者分析中，主要思路是用手工设计特征，使用机器学习的方法进行分析。周靖挑选出 100 个特征词汇，并统计出各个特征词汇在各章的词频，分别用 Bagging、Adaboost 和 Rotation Forest 方法预测各章回的作者类别，得出《红楼梦》前 80 回与后 40 回是由不同的人来完成的结论[5]。姜娜娜从虚词、长短句、词性标注和特有词四个主要特征入手，分别使用 SVM、Logistic 回归和 K-means 三种算法，也证实前 80 回与后 40 回的作者不是同一人的结论[6]。

由此可以看出，目前的红楼梦作者问题研究，主要方法是针对人工设计的特征进行统计分析或送入

基金项目：上海高原学科建设项目（10-17-303-004）。

作者简介：朱东旭（1996-），男，硕士研究生，主要研究方向：自然语言处理、机器学习、系统科学与复杂网络；严广乐（1957-），男，博士，教授，主要研究方向：系统科学与复杂网络、非线性动力学。

通讯作者：严广乐 Email：glyan2003@sina.cn

收稿日期：2020-05-29

分类器。这其中存在两个主要问题：

(1)人工选取的文本特征具有一定的主观性[7]；

(2)没有设置针对其它分界点以及其它作品的对照实验，因而可信度有待提高。

近年来，深度学习作为一种计算框架在许多研究领域中的迅速普及并取得了令人惊讶的成绩[8-10]。在自然语言处理、计算机视觉方面的表现，达到了与人类相同甚至超越人类的水平[11-12]。

为解决上述红楼梦作者研究中的问题，可利用深度神经网络的特征自动提取能力和数据拟合能力，对文本的风格特征进行自动提取并进行分类，用于不同作者分界线的确定。为此，本文提出了一种基于深度学习的文本风格分界点识别方法及相应的深度神经网络模型。

1　文本风格分界点识别

文本风格分界点是指：以此点为分界点前后两部分文本能获得最大风格差异的点。红楼梦作者问题本质上也是寻找前后文本差异最大的分界点，所以也属于文本风格分界点识别问题。

1.1　文本风格分界点识别方法框架

假设数据集 S 为 N 个无标签样本组成的集合。其中前 M 个样本属于第一类样本，后 $N-M$ 个样本属于第二类样本，即 M 是样本的分界点，但 M 的实际值未知。令 x 表示训练分类器的训练分界点，x^* 为测试准确率最高的训练分界点，将其命名为最优分界点；M 为样本集合的实际分界点；$A(x)$ 表示训练分界点为 x 时，分类器的测试准确率。由上述设定可知，当 $x=M$ 时，即训练分界点等于实际分界点时，分类器准确率为 $A(M)$。根据定性分析可知，由于训练数据中存在不同类型的样本被标记为相同标签的情况。所以，在宏观上分类器无法提取出将样本按标签准确划分的特征，从而无法得到好的泛化效果，测试准确度较低；其次，在训练过程中由于同一标签中存在的样本特征不同，所以其在梯度下降时存在一些相互抑制的梯度变化，降低了训练的效果。当 $x \neq M$ 的时 $A(x) < A(M)$，所以 $x^* = M$，其示意图 1 所示。其中，M 的实际值无法提前知道，其是求解的目标，运用 $x^* = M$ 这条性质可以确定出 M 的实际值。

1.2　分界点识别框架正确性验证

1.2.1　实验数据的选择

为了获取更加明显的实验效果且不失一般性，本文在 MNIST 随机选取了 1 000 张数字“1”与 1 000 张数字“8”组成实验的样本集合。其中样本集合的前 1 000 张为数字“1”，后 1 000 张为数字“8”。

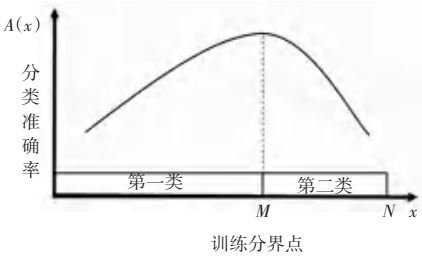

图 1　测试分类准确率随训练分界点的变化图

Fig. 1　Change of test classification accuracy with training boundary point

图 2　数字“1”与数字“8”部分样本

Fig. 2　Sample of number "1" and number "8"

1.2.2　实验数据生成

实验数据的生成应保证在不同训练分界点 x 处生成的数据集均需数据平衡且总数据量相近，从而尽可能减弱其它因素的干扰。数据生成步骤如下：

(1)设置训练分界点 x 的取值集合 X 为{600，650，700…1 400}。

(2)根据某一训练分界点 x，为所有样本打上标签。

(3)为防止数据不平衡，对数据集进行重采样。

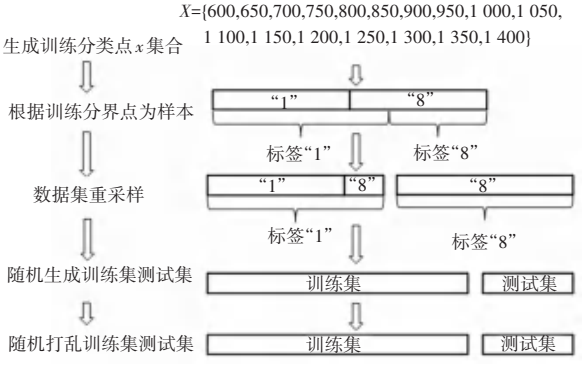

图 3　实验数据集生成步骤

Fig. 3　Steps of generating experimental data set

(4)在重采样后的样本集合中，随机选择 600 个样本作为测试集，剩余样本作为训练集。

(5)对训练集、测试集随机打乱顺序。

1.2.3 模型训练及结果分析

本实验采用三层二维卷积层、两层全连接层组成分类器,执行本次分类任务。具体结构如图 4 所示。

从图 5 可以看出,epochs 等于 30 时模型性能已经没有明显上升,所以可将训练 epochs 设定为 30。为减弱模型训练的不确定性,在每组训练集、测试集上,重复训练 15 次,获得 15 个分类器,分别测试其测试集准确度。

在不同训练分界点处,训练得出的模型在测试集上的准确度分布如图 6 所示。在图中可以明显看出,在训练分界点等于实际分界点时,模型的准确度最高;模型准确度随偏离程度增大不断下降,这与之前理论分析结果完全相同,证明了分界点识别框架的正确性。

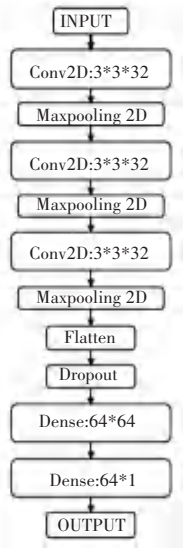

图 4 实验分类器具体结构

Fig. 4 Specific structure of experimental classifier

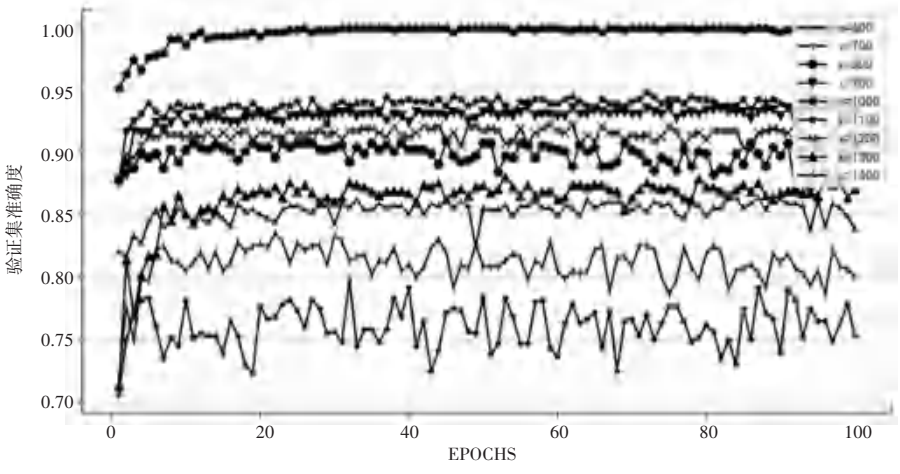

图 5 训练过程中验证集准确度变化图

Fig. 5 Change of validation set accuracy during training

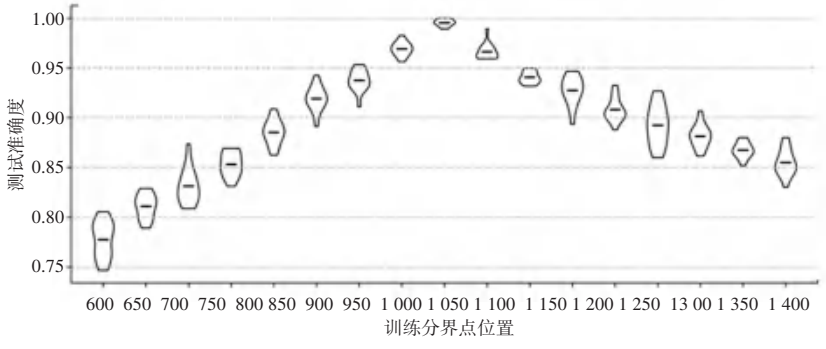

图 6 测试准确率随训练分界点变化图

Fig. 6 Change of test accuracy with training boundary point

2 红楼梦文本风格分界点的识别

2.1 数据收集与预处理

由于原始文本中包含标点、书名、人物名称和地名等内容,尤其是重要的人名和地名会随故事情节的发展而发生变化,给文本风格分类带来了很大的噪声干扰。为了尽可能去除噪声干扰,本文使用自

然语言处理中的相关技术对数据集进行预处理。根据句号、问号、叹号和省略号，将原始文本划分为句子，并去掉所有空格和符号；之后使用 jieba 分词工具进行分词。为防止剧情的发展对分类的影响，收集了文本中出现过的人物名称和地点名称，并将所有人物名称和地点名称分别用“人物名称”和“地点名称”代替。通过实验观察到，经过这样处理后结果现象更加明显。

2.2　文本词向量表示

本文利用 Word2Vec 工具[13]，对预处理后的文本进行训练，从而获取适合文本的词嵌入。Word2Vec 可以根据语义关系，将词条映射为 n 维向量，使具有相似语义的词条拥有相近的词嵌入，在很大程度上提升了算法对词的表征效果，方法简捷高效。

为解决稀疏词条的处理问题，本文忽略在红楼梦全文仅出现一次的词条。这样既保证了文章大部分词条得到保留，又降低了大量的稀疏词条对结果的影响。

在训练词向量时，另一个重要的问题是词嵌入维数 e 的选择，词嵌入维数太大或太小都会影响模型效果。经过实验，考虑到模型的性能与计算速度，最终确定词嵌入为维数 $e = 16$。使用本文 1-DconvLSTM 模型在不同词嵌入维数下的训练情况如图 7 所示。

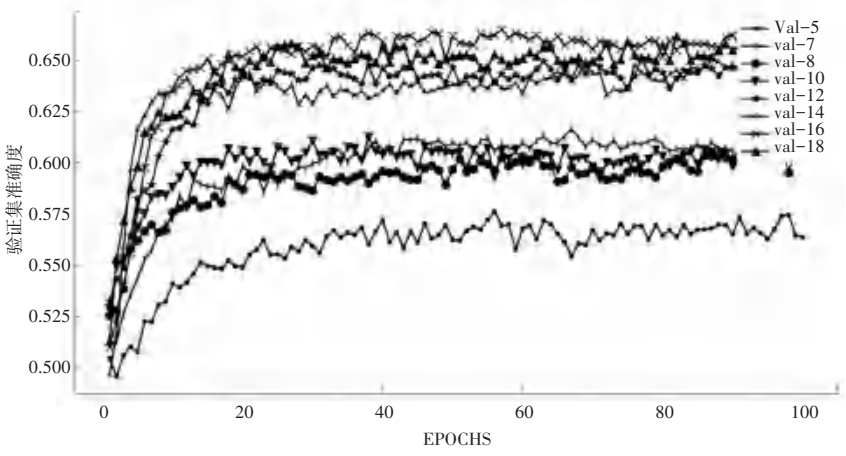

图 7　不同词嵌入维度下的模型的训练情况图

Fig. 7　Training situation of models with different word embedding dimensions

2.3　实验数据集

在生成数据集的过程中，主要面对 2 个问题：(1)数据穿越问题；(2)数据不平衡问题[14]。为解决上述问题，实验数据集生成步骤如下：

(1)将数据集使用词向量表示记为 D。

(2)将 D 按章节分为 120 份，每一章用 d_i 表示，并随机确定比例一定的章节用于测试。

(3)在 d_i 内部，将连续的 n 个词向量作为一个样本，即采样窗口大小为 n。

为防止出现因为训练分界点的设置而导致的数据不平衡，以及在不同训练分界点的样本总数量相差很大的问题，本文采用自适应调整采样步长的方法。采样步长设置为：假定当前的训练分界点为 C，则属于第一个作者章节的采样步长为 $s_1 = \frac{C}{60} \cdot s$，属于第二个作者章节的采样步长为 $s_2 = \frac{120 - C}{C} \cdot s_1$，其中 s 为基准采样步长。这样可以通过自动调节采样步长平衡两类数据的比例近似为 1∶1，并且在不同分界点处的样本总量近似相等，减弱了章节数量不同带来的影响。

(4)对每一个样本根据其所在的章节以及训练分界点设置标签。

(5)分别打乱训练集和测试集的样本顺序。

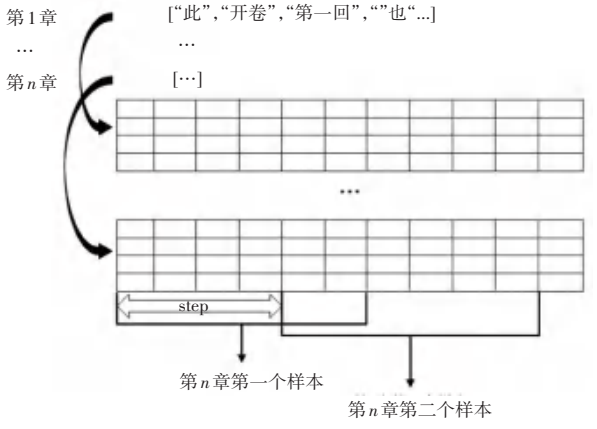

图 8　实验数据集生成中采样示意图

Fig. 8 Sampling diagram in experiment data set generation

2.4　作者风格识别模型

由于输入到模型的数据已经由文本变为了长度

为窗口长度 n，宽度为词嵌入维度 e 的二维张量。输入可以看作序列长度为 n 的 e 维的时间序列，词嵌入的不同维度可以看成不同的数据通道。所以，使用 1-D 卷积来提取不同通道局部特征，用 LSTM 提取序列整体特征。

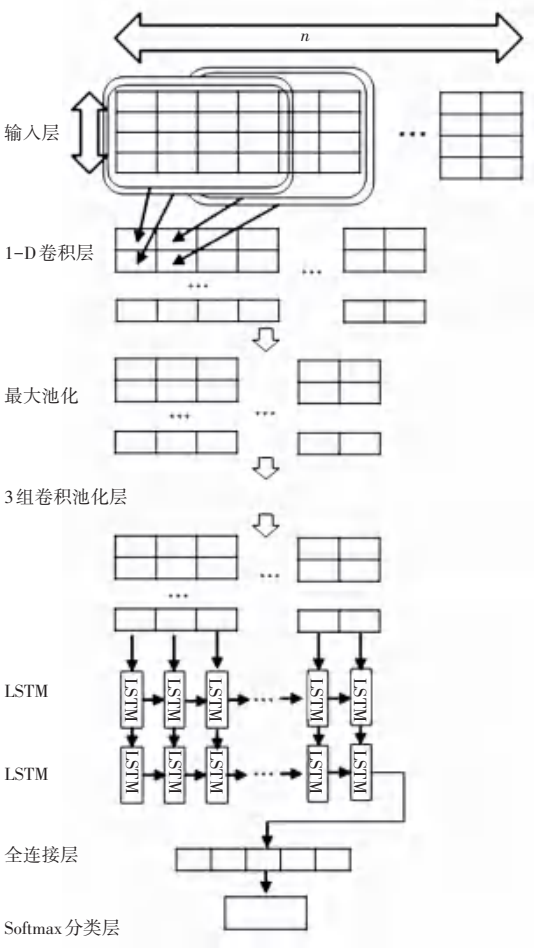

图 9 1-DconvLSTM 模型结构图

Fig. 9 Structure of1-DconvLSTM model

本文构建的 1-DconvLSTM 模型由输入层、1-D 卷积层、最大池化层、LSTM 层和全连接层分类层构成，并使用批标准化技术，具体结构如图 9 所示。

利用该模型可获得较好的作者风格识别效果，模型各部分作用如下：

(1)输入层。该层为模型的输入部分，输入数据为经过采样后长度为 n、宽度为 e 的矩阵。其可以表示为如下形式：

$$Sen = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{11} & \cdots & s_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{e1} & s_{e2} & \cdots & s_{en} \end{bmatrix}, \quad (1)$$

其中，Sen 中的第 i 列代表样本第 i 个词所对应的 e 维词向量。

(2)1-D 卷积层。本层的作用是进行局部特征的抽取。在 1-D 卷积层中，卷积核对输入的多维时间序列进行卷积操作，提取时间序列中的模式信息，然后经过非线性激活函数产生本层的输出，根据文献[15]，其数学模型可以描述为：

$$x_{ik}^{l+1} = f(W_k^l \cdot x_{i:i+h-1}^l + b_k^l). \quad (2)$$

式中，x_{ik}^{l+1} 为 $l+1$ 层第 k 个通道中第 i 个输入，W_k^l 为 l 层第 k 个通道的卷积核，b_k^l 为 l 层第 k 个通道的偏置项，f 为激活函数。

根据文献[16]的研究表明，在保证参数量不变的情况下，使用多层小卷积核的效果优于使用大卷卷积核效果。所以，在此使用了 4 层核，大小为 3 的卷积层。

(3)最大池化层。最大池化算子用来提取输入特征中的局部最大值，有效降低了可训练参数的数量，提高了模型训练的速度，并在一定程度上提升了模型的泛化能力。

(4)批标准化层。本操作的目的是对输入到各层的数据进行标准化，使各层输入的分布近似保持不变。根据文献[17]，批标准化操作可以有效提升模型的泛化能力，并且加快模型的训练速度，缓解了梯度消失问题。

(5)LSTM 层。长短期记忆网络 LSTM 是一种特殊的 RNN 网络，其通过在神经单元中增加门控机制，克服了 RNN 的长期记忆问题，可以有效学习到长期依赖关系。

文献[18]的研究表明，在处理时间序列时，使用两个及以上的递归层，对模型性能的提高是有帮助的。所以在本文的模型中使用了两个连续的 LSTM 层，以提高模型性能。

(6)全连接层。经过前面多层的处理，最终得出的特征向量会作为本层的输入，经过全连接层的非线性映射后送入分类器，进一步提升模型的非线性拟合能力。

3 实验与结果分析

红楼梦由于历史问题，流传下来的版本众多，但其主要分为两大版本系统，一个是仅流传前八十回的脂评系统；另一个是经程伟元、高鄂收集整理的一百二十回的程高本系统。关于红楼梦作者的争论主要存在于程高本的后 40 回是否是由原作者所著。考虑到程高本前 80 回可能也对原文做了一定程度的修改，从而影响到文本风格的判断的准确性，所以本文在使用程高本作为鉴别样本的基础上，将岳麓书社整理的 80 回脂评本与程高本的后 40 回拼接构成第二个鉴别样本。在两个样本上进行独立的实验

和判断，以提高实验的准确性与可信度。

3.1　实验数据集

本文生成的实验数据集如表 1 所示。其中Ⅰ类代表样本位于训练分界点之前，Ⅱ类样本位于训练分界点之后。从表 1 中可以看出，不同分界点的数据总数相似且数据平衡，这样可以有效降低其它因素对模型性能的影响，从而保证了分界点识别的准确性和可靠性。

表 1　数据集样本分布表

Tab. 1　Data set sample distribution

训练分界点	训练集样本数				训练集样本数			
	第一鉴别样本		第二鉴别样本		第一鉴别样本		第二鉴别样本	
	Ⅰ类	Ⅱ类	Ⅰ类	Ⅱ类	Ⅰ类	Ⅱ类	Ⅰ类	Ⅱ类
50	10 444	12 335	10 523	13 022	3 168	4 131	3 233	3 694
55	10 460	12 063	10 853	12 360	3 552	4 249	3 172	4 320
60	10 196	11 580	10 072	11 425	3 270	3 411	3 482	3 858
65	10 621	12 162	10 668	11 875	3 464	3 329	3 552	3 899
70	10 941	11 858	11 092	12 283	3 395	3 407	3 359	3 294
75	10 739	12 463	11 097	12 638	3 896	3 880	3 724	3 942
80	10 556	10 886	10 646	10 855	3 561	3 578	3 738	3 637
85	10 844	11 834	10 662	12 300	3 503	3 872	3 958	3 413

3.2　评价指标

常见的二分类评价指标有精确度、召回率、F1 值和准确率。由于本实验中没有明显的正负类之分，并且数据分布是平衡的，所以选择准确度 A 作为评价指标。

准确度 A 表示对测试集进行分类后，所有预测正确的样本占总样本的比例，即：

$$A = \frac{TP + TN}{TP + FP + TN + FN}.$$

表 2　混淆矩阵表

Tab. 2　Confusion matrix table

		真实值	
		1	0
预测值	1	TP	FP
	0	FN	TN

3.3　模型训练与结果分析

1-DconvLSTM 网络模型训练的具体设置见表 3。对于每一个训练分界点，重复训练 15 个模型，以降低偶然因素。

两个鉴别样本在不同训练分界点处训练模型的测试准确率分布如图 10 和图 11 所示。从两幅图中可以看出：

(1)在训练分界点为 80 回时，两个鉴别样本模型的准确率都取得了最高值以及最高的平均值。由此根据前面的结论可以推断出，在两个鉴别样本中，均是以 80 回为分界点时，前后的风格差异最大，即 80 回是它们的实际风格分界点。

(2)两幅结果图之间存在细微的差别。第二幅结果图更接近图 6 在理想情况下得到的结果，而第一幅结果图却出现了一定程度的波动。这说明虽然两个鉴别样本在 80 前后都存在明显的风格变化，但是第二鉴别样本在 80 回处前后的差别相对于第一鉴别样本来说更加明显，从而印证了程高本对前 80 回进行了一定程度改动的观点。

表 3　模型训练设置表

Tab. 3　Model training setting table

参数/函数名称	参数值/函数值
激活函数	RELU
代价函数	二元交叉熵损失函数
优化方法	Adam
加速方法	批标准化技术
训练迭代次数	50
词向量维数	16
采样窗口大小	40
采样步长	自适应，基准步长为 15
重复训练模型个数	15

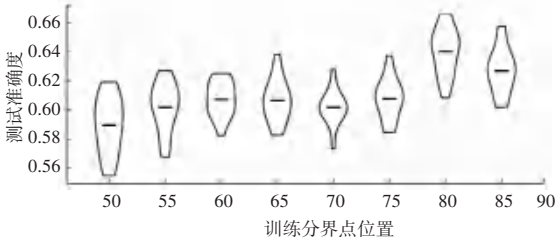

图 10　第一鉴别样本测试准确度随训练分界点变化图

Fig. 10　Test accuracy of the first identification sample changes with the training boundary point

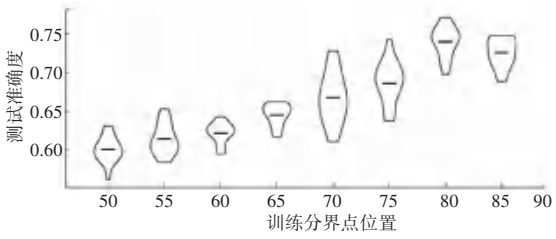

图 11 第二鉴别样本测试准确度随训练分界点变化图

Fig. 11 Test accuracy of the second identification sample changes with the training boundary point

3.4 对比试验

为了进一步提高结论的可信度,对确定只有单一作者的著作《水浒传》使用本文的方法进行分析。

在经过相同的实验步骤处理后,《水浒传》得出的结果如图 12 所示。

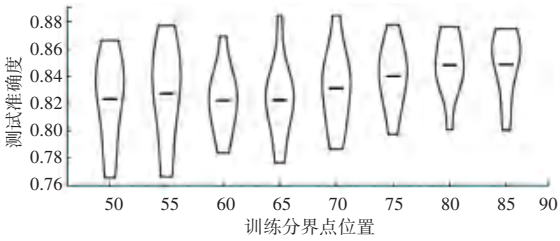

图 12 水浒传测试准确度随训练分界点变化图

Fig. 12 Water Margintest accuracy changes with training boundary point

由图中可以看出,单一作者的著作在不同训练分界点处的测试准确度变化不大,而且没有出现明显的变化趋势;而在红楼梦的两个鉴别样本中却都出现了在 80 回处明显的准确度提升现象,且在第二鉴别样本的结果图中出现了明显的趋势。通过对比,进一步说明红楼梦在 80 回前后确实出现了文本风格的明显变化,并且有较高的可信度。

4 结束语

在传统的红楼梦作者问题研究中,使用手工设计的特征作为文本特征有一定的主观性,并且缺少横向与纵向的对照实验,使结果不够可信。故本文针对以上问题,提出一种基于深度学习的文本风格分界点识别方法以及一种用于文本分类的深度模型 1-DconvLSTM 模型。实验结果表明,红楼梦的文本风格分界点在 80 章处,强有力的支持了红楼梦 80 章前后不是同一作者所著的结论。

但是本方法需要在不同训练分界点反复训练模型以降低随机因素的影响,所需的计算量较大,计算较为耗时。所以接下来将尝试通过提出更加轻量化的模型、更高效的计算方式来降低计算时间。

参考文献

[1] 陈大康. 从数理语言学看后四十回的作者——与陈炳藻先生商榷 [J]. 红楼梦学刊, 1987(1): 293-318.

[2] 李贤平.《红楼梦》成书新说 [J]. 复旦学报:社会科学版, 1987(5): 3-16.

[3] 赵冈, 钟毅. 红楼梦新探[M]. 文化艺术出版社, 1991.

[4] 韦博成.《红楼梦》前 80 回和后 40 回某些文风差异的统计分析(两个独立二项总体等价性检验的一个应用) [J].《应用概率统计》. 2009, 25(4): 441-448.

[5] 周靖. 基于机器学习的《红楼梦》作者问题研究 [D]. 昆明: 云南大学, 2018.

[6] 姜娜娜. 基于机器学习的《红楼梦》作者研究 [D]. 杭州: 浙江大学, 2018.

[7] 叶雷. 基于计量文体特征聚类的《红楼梦》作者分析 [J].《红楼梦学刊》, 2016(5): 312-324.

[8] JIAO L, ZHANG F, LIU F, et al. A Survey of Deep Learning-Based Object Detection [J/OL]. IEEE Access, 2019(7): 128837-128868. (2019-9-05) [2019-12-30]. https://doi.org/10.1109/ACCESS.2019.2939201.

[9] YOUNG T, HAZARIKA D, PORIA S, et al. Recent Trends in Deep Learning Based Natural Language Processing [J]. IEEE Computational Intelligence Magazine, 2018(3): 55-75.

[10] BAI S, KOLTER Z J.KOLTUN V. An Empirical Evaluation of Generic Convolutional and Recurrent [OL]. arXiv preprint. arXiv: 1803.01271v2. (2019.04.19) [2019.12.30]. https://arxiv.xilesou.top/abs/1803.01271v2.

[11] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [OL]. arXiv preprint. arXiv: 1810.04805v2. (2019.05.28) [2019.12.30].

[12] WANG X, HUANG Q, CELIKYILMAZ A, et al. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Science, 2019: 6629-6638.

[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Proc of the 27th Annual Conference on Neural Information Processing Systems. New York: Curran Associates 2013, 2013: 3111-3119.

[14] 叶枫,丁锋.不平衡数据分类研究及其应用[J].计算机应用与软件,2018,35(01):132-136,205.

[15] KIM Y. Convolutional neural networks for sentence classification [C]// Proc of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.

[16] HE K, SUN J. Convolutional neural networks at constrained time cost [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Science, 2015: 5353-5360.

[17] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// Proc of the 32nd International Conference on Machine Learning. 2015: 448-456.

[18] KARPATHY A, JOHNSON J, LI F F. Visualizing and understanding recurrent networks [OL]. arXiv preprint. arXiv: 1506.02078, 2015. (2015.11.28) [2019.12.17]. https://arxiv.xilesou.top/abs/1506.02078v2.