第 57 卷 第 6 期 厦门大学学报(自然科学版) Vol. 57 No. 6
2018 年 11 月 Journal of Xiamen University (Natural Science) Nov. 2018
doi:10.6043/j.issn.0438-0479.201804012

基于降噪自编码器特征学习的作者识别及其在《西游记》诗词上的应用

范亚超,罗天健,周昌乐*
(厦门大学 信息科学与技术学院,福建省类脑计算技术及应用重点实验室,福建 厦门 361005)

摘要:由于作者归属问题较为复杂,采用传统自然语言处理模型难以完成作者识别.为了深入挖掘作者归属问题,首先采用降噪自编码器深度模型提取文本结构特征,再采用支持向量机分类器完成作者识别.模型的优势在于能够考虑未知文本特征的噪声多样性和复杂性,且能够重构添加噪声的原始文本输入.将该方法应用于吴承恩、王廷陈、薛蕙等人的诗词作者识别,识别准确率最高为 78.2%,验证了该方法的有效性,进一步将该方法应用于《西游记》诗词作者识别.

关键词:降噪自编码器;编码特征;作者识别
中图分类号:TP 391.1 **文献标志码**:A **文章编号**:0438-0479(2018)06-0884-06

作者识别是从给定的一系列作者中确定未知文本的作者.在机器学习角度下,该问题可以转化为文本风格的多分类问题,即将每个作者的文本作为一个类别,通过风格特征进行分类.作者风格特征是指在同一作者文本中稳定出现且特有的模式.

文本风格和作者识别研究可以追溯到 19 世纪 60 年代,Mendenhall[1]以不同长度词汇的频率分布作为不同作者的特征形式.在 20 世纪中叶,许多统计学方法被用来衡量写作风格,包括齐普夫(Zipf)分布和 Yule 相关系数.Mosteller 等[2]最早以常用功能词为文本风格特征应用贝叶斯统计方法分析进行现代作者识别.

目前已提出很多体现文本风格的可量化的语言结构特征[3],包括词汇丰富度[4]、虚词使用频率[5-7]、高频词[6]、成词率[3]、词共现网络[8]、网络嵌入[9]、基于词的多元文法[5,7]特征,而对中文诗词和文言文的作者识别目前研究较少.施建军[6]以 44 个文言虚字使用频率为特征,采用支持向量机模型对《红楼梦》作者进行研究.肖天久等[7]以独有词、虚词及词类的多元文法进行聚类,通过计算相似度对《红楼梦》作者进行分析.易勇等[10]以诗词向量空间模型为特征,提出基于机器学习的朴素贝叶斯方法对李白和杜甫的作品的作者进行判别.

如今,深度学习模型已经成功应用于不同自然语言处理任务,在各项识别任务上优于现有方法,为作者识别问题的研究提供了新方法.在作者识别研究中传统机器学习方法虽然取得了较好的结果,但是仍然存在缺陷,例如在特征工程中较难提取到能够很好表征不同作者写作风格的特征,而特征的选择将直接影响分类结果.相较于基于统计的特征提取方法[10],对相同诗词流派的作者识别问题,神经网络在训练过程中能够充分学习到不同作者的用词分布特征和结构特征,降噪自编码器(DAE)模型[11]是神经网络中常用的文本特征提取方法,该模型在经过编码和解码后能够同时将这些特征进行融合,从少量的数据中获取更高级特征,能够得到更好的识别效果.因此,本研究将采用 DAE 模型提取编码特征,通过无监督学习预训练模型得到高阶特征向量,从量化计算角度分析作者归属问题.

1 整体分析流程和实现

本研究的分析流程如图 1 所示,首先进行特征提

收稿日期:2018-04-23 录用日期:2018-10-04
基金项目:国家自然科学基金(61673322,61573294)
*通信作者:dozero@xmu.edu.cn
引文格式:范亚超,罗天健,周昌乐.基于降噪自编码器特征学习的作者识别及其在《西游记》诗词上的应用[J].厦门大学学报(自然科学版),2018,57(6):884-889.
Citation:FAN Y C,LUO T J,ZHOU C L. Author identification based on denoising autoencoder and it's application in" Journey to the West"[J]. J Xiamen Univ Nat Sci,2018,57(6):884-889.(in Chinese)

取,特征提取是从文本中提取能够代表文本风格的特征,并将这些特征量化为特征向量用以进行计算. 目前已经提出了很多文本风格特征,本研究采用高频字和 53 个文言虚词使用频率,这 2 种特征在前人的工作中采用[5-7],且被证明在作者识别任务中具有明显效果[12]. 高频字特征可以反映作者在写诗词时的常用字词,从侧面反映作者的用字灵活程度;虚词主要包括介词、连接词、助词、语气词及副词等,与作品主题无关,是作者在长期写作过程中形成的习惯用词方式,因此在文本风格分析中多被使用.

第二步进行特征选择,特征选择是用于去除冗余特征和保留那些能够高度表征文本风格的特征,以达到避免过度拟合、提高准确率和减少训练时间的效果. 本研究选取基于使用频率的方法,即选择在语料库中出现次数最多的 N 个特征.

特征确定之后,对样本进行量化得到所有样本的特征向量,将其作为模型输入进行特征学习,再采用不同的解码器预训练 DAE,然后将选择的频率特征输入微调后的分类神经网络模型,进行降维提取编码特征,最后比较编码特征与现有特征的分类准确率.

最后进行分类训练和测试,支持向量机在解决小样本、高维特征模式识别中表现出很多特有优势,且算法简单,有很好的鲁棒性[5-6,13],因此在很多研究中被广泛应用. 本研究采用基于径向基核函数的支持向量机分类器.

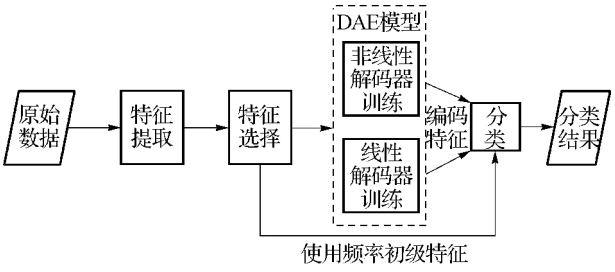


图 1　作者识别流程

Fig. 1　The processing of the authorship identification

2　DAE

2.1　DAE 无监督训练过程

由于 DAE 在输入值中加入人工噪声,使其从带噪声的数据中学习到目标数据的特征,更能抗噪声干扰,所以 DAE 的泛化能力比自编码器更好. 该模型由两部分组成:编码器和解码器,编码器在输入层和隐含层之间,而解码器位于隐含层和输出层之间. 下面将介绍 DAE 的训练过程,如图 2 所示.

1) 为了提升模型的泛化能力,添加噪声到输入 $x \in \mathbf{R}^d$, d 是输入向量维度,将 x 转换为 $\tilde{x}$, 使其更符合实际. 以 2 种方式对模型输入添加噪声,即服从参数为 p 的伯努利分布噪声和高斯噪声.

2) 用非线性转换方程 f_{en} 将 $\tilde{x}$ 编码到隐含层输出 $y=(y_i)$,

$$y_i = f_{\text{en}}(t_i),$$

其中: $y \in \mathbf{R}^h$; f_{en} 为编码方程,本文中采用 sigmoid 函数; t_i 为 t 的第 i 个元素, $t = \boldsymbol{W}_{\text{en}}\tilde{x} + \boldsymbol{b}_{\text{en}}$, $\boldsymbol{W}_{\text{en}} \in \mathbf{R}^{h\times d}$ 为编码层的权重矩阵, $\boldsymbol{b}_{\text{en}}$ 为偏置量.

3) 隐含层输出向量 y 通过解码层重构输出值 $z=(z_i)$,

$$z_i = f_{\text{de}}(t_i'),$$

其中: $z \in \mathbf{R}^d$; f_{de} 为解码方程,本研究采用线性解码函数和 sigmoid 函数进行对比; t_i' 为 t' 的第 i 个元素, $t' = \boldsymbol{W}_{\text{de}}y + \boldsymbol{b}_{\text{de}}$, $\boldsymbol{W}_{\text{de}} \in \mathbf{R}^{d\times h}$ 为解码层的权重矩阵,与编码层权重共享,即 $\boldsymbol{W}_{\text{de}} = \boldsymbol{W}_{\text{en}}^{\mathrm{T}}$, $\boldsymbol{b}_{\text{de}}$ 为偏置量.

4) DAE 的目标是提取优化特征,它能够从 $\tilde{x}$ 中鲁棒地得到,并且可以重构输出值 z, 使其尽可能恢复出真实输入值 x. 计算过程中通过损失函数计算重构误差,再利用反向传播优化参数. 当采用 sigmoid 解码器时,损失函数采用交叉熵,

$$\text{cost} = -\frac{1}{d}\sum_{i=1}^{d} x_i\log(z_i) + (1-x_i)\log(1-z_i);$$

当采用线性解码器时,损失函数采用均方误差,

$$\text{cost} = \frac{1}{2d}\|\boldsymbol{x}-\boldsymbol{z}\|^2.$$

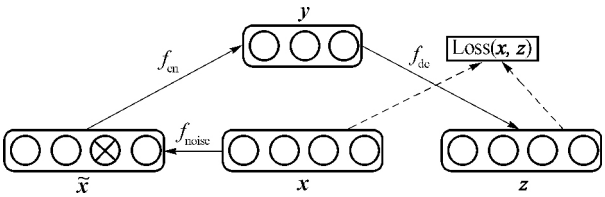


图 2　DAE 结构

Fig. 2　The architecture of DAE

2.2　有监督微调训练

经过无监督训练后,隐含层学习获得了能够很好表征输入数据的编码特征,但是 DAE 模型并不能实现分类,需要在分类器的训练下实现数据分类. 因此,通过添加逻辑回归分类器对预训练的参数进行微调,如图 3 所示,训练过程如下:

1) 在 DAE 的编码层后添加逻辑回归层.

2) 用预训练模型编码层权重初始化该神经网络对应层权重.

3) 用带标签的数据微调神经网络参数,避免出现局部最优.

$\boldsymbol{T} = f_{\sup}(\boldsymbol{q})$,

其中:$\boldsymbol{T}$ 为全连接层的输入;$f_{\sup}$ 为线性激活函数;$\boldsymbol{q} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{b}, \boldsymbol{W} \in \mathbf{R}^{h\times c}$ 为全连接层权重矩阵,$\boldsymbol{b}$ 为偏置量.根据 $\boldsymbol{T}$ 可求得预测分类概率矩阵 $\boldsymbol{p}_{\text{pred}}$,

$$p_{\text{pred}_i} = \frac{e^{T_i}}{\sum_{i=1}^{c} e^{T_i}},$$

其中 c 为最终分类个数.

经过无监督预训练和有监督微调后,将每个输入值 x 输入分类神经网络提取特征,最终作为编码特征.

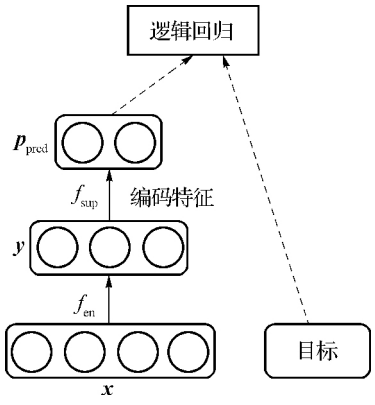


图 3 分类神经网络微调

Fig. 3 Fine-tuning of a neural network for classification

3 实 验

为避免因为文本体裁不同造成的影响,本研究首先选取《吴承恩诗文集》[14]中部分诗词及与吴承恩同一时期的其他三位代表性的人物诗集作品包括王廷陈《梦泽集》[15]、薛蕙《考功集》[16]、杨慎《升庵集》[17]对模型进行训练和测试,用以验证模型的有效性;再对世德堂本章回小说《西游记》中诗词部分(去除描写人物和打斗诗词)共 388 首进行作者识别,为了使样本均衡,取每个作者作品集中部分诗词,诗词样本数量如表 1 所示.本研究选择虚词如下:焉、兮、于、因、在、乎、以、虽、与、使、未、又、况、犹、的、是、至、几、莫、了、便、勿、惟、皆、自、空、却、着、且、不、方、如、若、即、为、矣、倘、纵、己、曾、则、乃、其、而、何、尚、所、之、者、也、仍、自、非[18-20].

表 1 样本统计信息

Tab. 1 Sample statistics

作品	样本数量	用字数
《吴承恩诗文集》	644	15 808
《梦泽集》	670	17 074
《考功集》	637	13 280
《升庵集》	637	14 012
《西游记》	388	22 728

语料库确定之后,统计所有用字,并以使用次数最多的方法选择高频字和上述选用的虚字作为特征,再用词袋模型构建每个样本的特征向量 x,最终所有样本的特征向量作为训练该问题的数据集.采用 5 折交叉验证方法,将数据集随机分为 5 份,其中每份轮流作为测试集用以评估分类效果;其余 4 份为训练集;每个组合独立训练,最终对 5 次测试结果取平均值即为最终结果.实验训练步骤为首先采用 DAE 模型把每个样本的特征向量 x 作为输入,z 为经过模型还原的样本特征向量,训练的目标是使得 z 尽可能地接近原样本特征向量 x,最终获得降维后能够表征作者文本风格的低维向量 y,DAE 模型训练好后再经过分类神经网络微调 DAE 模型编码层参数.所有训练结束后把全部数据集使用分类神经网络模型映射为编码特征,然后比较现有特征和编码特征的分类准确率.表 2 列出了 DAE 网络结构的相关参数.

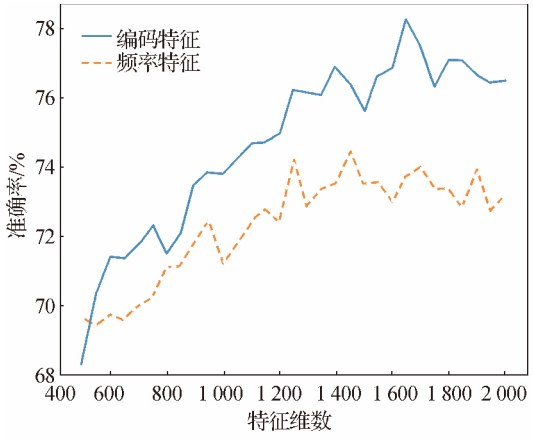
表 2 神经网络的超参数

Tab. 2 List of hyperparameters for the neural network

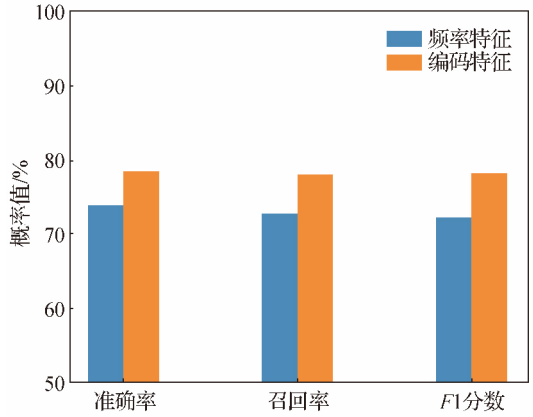
隐含层节点数	预训练迭代次数	预训练学习率	微调学习率	噪声级别
500	50	0.001	0.01	0.3

3.1 编码特征和频率特征的结果比较

本节将比较编码特征与频率特征的分类准确率.分别提取 4 位诗人诗词作品中的虚字和高频字使用频率特征,DAE 解码层采用线性激活函数,然后用本研究提出的方法提取编码特征,通过支持向量机方法比较频率特征和编码特征的分类指标,如图 4 所示.图 4(a)显示当采用不同特征维数时,采用两种特征的分类准确率,可以看出随特征维数的增大准确率整体上呈上升趋势,且编码特征的准确率明显优于使用频率特征.图 4(b)为当特征维数为 1 650 时,采用 2 种特



(a) 2 种特征分类准确率比较



(b) 当频率特征维数为1 650时，两种特征分类结果比较

图 4　编码特征和频率特征结果比较

Fig. 4　Comparison of the results of code feature and frequency feature

征对吴承恩进行分析得出分类指标，包括精确率、召回率和 $F1$ 分数，结果显示编码特征在各项指标均优于频率特征. 综合分析表明采用本研究提出的方法得到的编码特征效果优于频率特征. 接下来考虑不同 DAE 超参数设置对结果的影响.

3.2　不同解码器的结果比较

本研究通过在 DAE 解码层采用线性解码器和非线性 sigmoid 解码器比较两者对准确率的影响. 如图 5 所示，当特征维数较小时，两者准确率相差不大，随着特征维数增加，线性解码器对结果的提升越来越明显. 特征维数为 1 400 时，非线性 sigmoid 解码器的准确率达到最大值，之后随着特征维数增加准确率逐渐趋于稳定. 当特征维数为 1 650 时，线性解码器的准确率达到最大值. 因此当采用线性解码器时，可以对更高维数的特征进行处理，且准确率更高.

3.3　不同隐含层节点数的结果比较

DAE 中隐含层节点数为编码特征的维数，可以提取输入特征的更具代表性的特征，同时可以降低输入特征的维数，减少冗余特征，但降噪自编码器同时能够考虑到特征之间的非线性关系. 如图 6 所示，当隐含层节点数太少时，不能很好地表征输入特征，导致模型数据重构能力降低，同时也会降低最终的识别准确率；当隐含层节点数过多时，会导致提取到的特征同时也学习到了噪声的特性，降低编码特征的抗噪性能，预测准确率下降.

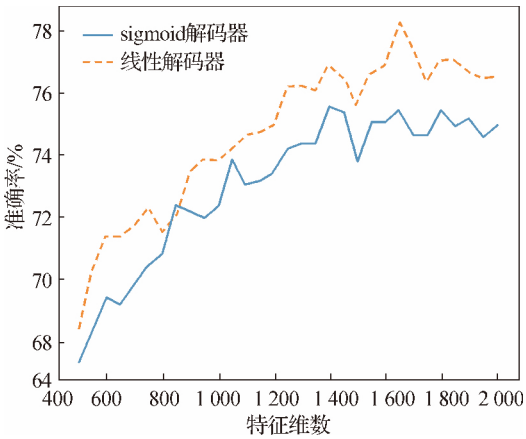


图 5　线性解码器和非线性解码器分类结果比较

Fig. 5　Comparison of the results of linear decoder and nonlinear decoder

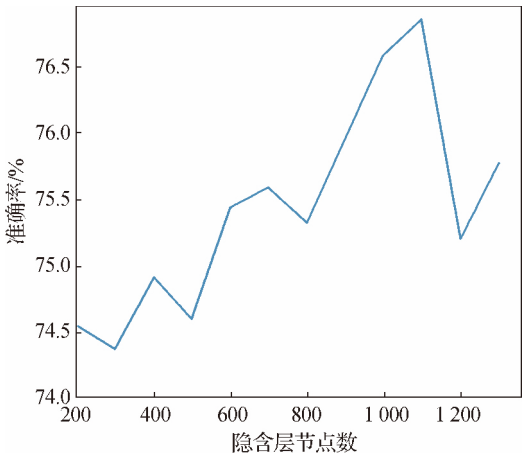


图 6　当特征维数为 1 650 时，比较采用不同隐含层节点数的准确率

Fig. 6　The accuracy in different nodes of hidden layer when feature dimension is 1 650

4　《西游记》中诗词预测结果讨论

《西游记》作者问题从小说问世到现在争议不断，研究者主要围绕文献记载和小说内容进行分析，集中

讨论作者是否为吴承恩,同时也提出了其他怀疑作者.杜贵晨等[21]对该问题做了详细描述,并指出根据近 20 年国内外研究结果,吴承恩说已经很难使学术界达成共识.

由于吴承恩遗留的小说作品仅有《西游记》,其他都是不同体裁的短篇文章,无法为本研究提供充足的训练预料,因此本研究仅对《西游记》中的诗词进行了作者判别.通过第 3 节分析,选取准确率最高的 3 种训练模型参数设置,对应 3 种特征维数分别为 1 650,1 700,1 800,解码器均为线性激活函数,编码特征维数隐含节点数均为 1 100.利用《吴承恩诗文集》的训练模型对《西游记》诗词进行预测,同样采用五折交叉验证,每次训练后均给出预测结果,最终每首诗对应 15 个预测结果,计算每首诗属于吴承恩的概率值,给出了《西游记》中诗词作者属于吴承恩的概率分布,如图 7 所示.

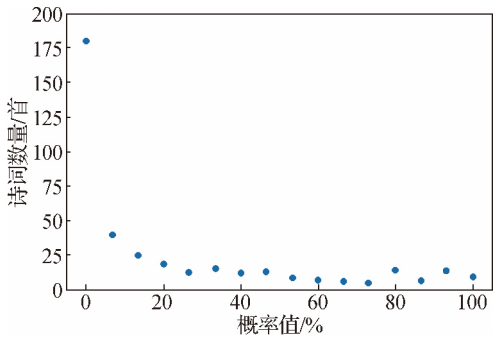


图 7 吴承恩所作诗词概率分布

Fig. 7 The probability distribution of the poems written by Wu Chengen

从图 7 可以看出,其中有 180 首预测概率为 0,说明这些诗词没有出现与《吴承恩诗文集》中用到高频词和虚词;预测概率大于 0 小于 50%的诗词有 136 首;大于 50%的诗词有 72 首,占选出的《西游记》诗词总数为 18.6%,其中预测概率为 90%以上的诗词数量仅有 27 首.从分析结果可以看出,从诗词角度,《西游记》中诗词与吴承恩的诗词存在相似之处,但数量不多.

同时,通过提取神经网络中分类权重较高的特征发现,西游记中较多出现数量词、景色相关字如“一”、“千”、“万”、“花”、树木等;吴承恩诗词中多出现与自然现象相关的字如“凤”、“云”、“日”、“月”等.从用词习惯可以看出不同作者在写诗词时寄托的事物和表现程度所采用的方式各有不同.

本研究同时通过每回目对应的诗词预测概率求和除以诗词总数,从诗词角度出发得出《西游记》各回目中诗词作者是吴承恩的相关概率,如图 8 所示.图中,X 轴表示回目序号,Y 轴表示各回目中预测为吴承恩诗词数量与对应回目所选取诗词数量之比,Z 轴表示预测诗词作者为吴承恩的概率.从图中可以看出概率较高的回目集中在开始回目和最后几个回目,同时预测的诗词数量在这些回目中所占比例也较高.从回目与相关概率的关系中可以看出,吴承恩与开头和结尾回目相关概率较大,但回目之间差别较大;与中间部分回目的相关概率较小,但回目之间差别较小;同时存在回目间隔相关,即全书间断性出现与吴承恩没有相关性回目.通过上述分析表明,在《西游记》诗词创作方面吴承恩的写作风格贯穿整部小说,但整体相关概率较小.

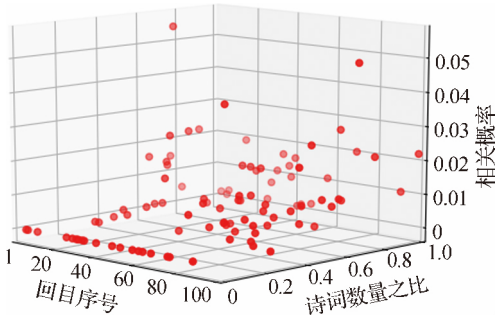


图 8 各回目与吴承恩的相关概率

Fig. 8 The relevant probability of each chapter and Wu Chengen

5 结 论

本研究提出了基于 DAE 特征学习的作者识别方法,采用 DAE 深度模型融合文本结构特征和用词特征提取到作者的用词之间的网络权重特征,再采用支持向量机分类进行作者识别,将该方法应用于吴承恩、王廷陈、薛蕙等同一时期诗人的诗词作者识别,识别准确率最高为 78.2%,验证了该方法的有效性.除此之外,本研究还对比了线性解码器和 sigmoid 解码器对分类准确率的影响,结果表明线性解码器较解码器更有优势.对比了 DAE 隐含层不同节点数对结果的影响,分析表明合适的隐含层节点数选择可以提升模型的鲁棒性.最后,通过本研究提出的方法对《西游记》中的诗词作者归属问题进行了实验,给出了《西游记》中诗词作者属于吴承恩的概率分布和各回目与吴承恩的相关概率.

今后的工作还需要在高频字和虚字等特征基础上,构建更完备的多维特征,并结合文章部分提出跨文本体裁模型综合分析,给出更全面的作者归属论据.

参考文献:

[1] MENDENHALL T C. The characteristic curve of composition[J]. Science,1887,9(S214):237-246.

[2] MOSTELLER F, WALLACE D L. Inference and disputed authorship: the federalist [J]. Revue De L Institut International De Statistique,1964,22(1):353.

[3] 肖天久,刘颖. 基于聚类和分类的金庸与古龙小说风格分析[J]. 中文信息学报,2015,29(5):167-177.

[4] STAMATATOS E,FAKOTAKIS N,KOKKINAKIS G. Computer-based authorship attribution without lexical measures[J]. Computers & the Humanities,2001,35(2):193-214.

[5] MOHSEN A M,EL-MAKKY N M,GHANEM N. Author identification using deep learning[C]// IEEE International Conference on Machine Learning and Applications. Anaheim:IEEE,2017:898-903.

[6] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊,2011(5):35-52.

[7] 肖天久,刘颖.《红楼梦》词和 N 元文法分析[J]. 现代图书情报技术,2015,31(4):50-57.

[8] 李晓军,刘怀亮,杜坤. 一种基于复杂网络模型的作者身份识别方法[J]. 图书情报工作,2015(18):102-107.

[9] CHEN T, SUN Y. Task-guided and path-augmented heterogeneous network embedding for author identification [C] // Tenth ACM International Conference on Web Search and Data Mining. New York:ACM,2017:295-304.

[10] 易勇,郑艳,何中市,等. 基于机器学习的古典诗词作者的判别研究[J]. 心智与计算,2007(3):359-364.

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

[12] 刘颖,肖天久.《红楼梦》计量风格学研究[J]. 红楼梦学刊,2014(4):260-281.

[13] EDER M. Does size matter? Authorship attribution, small samples, big problem [J]. Digital Scholarship in the Humanities,2015,30(2):167-182.

[14] 吴承恩. 吴承恩诗文集笺校[M]. 上海:上海古籍出版社,1991:1-357.

[15] 王廷陈. 梦泽集[M]. 台北:台湾商务印书馆,1969:5-368.

[16] 薛蕙. 考功集[M]. 上海:上海古籍出版社,1993:15-323.

[17] 杨慎. 升庵集[M]. 上海:上海古籍出版社,1993:212-530.

[18] 王力. 汉语语法纲要[M]. 北京:中华书局出版社,2015:409-418.

[19] 谢晓晖.《西游记》虚词"着"的词义探析[J]. 湖南第一师范学报,2004,4(4):74-76.

[20] 杨载武.《西游记》虚词"却"词义探[J]. 贵州师范学院学报,1994(1):28-34.

[21] 杜贵晨,王艳. 四百年《西游记》作者问题论争综述[J]. 泰山学院学报,2006,28(4):19-25.

Author Identification Based on Denoising Autoencoder and It's Application in "Journey to the West"

FAN Yachao, LUO Tianjian, ZHOU Changle*

(Fujian Keylab of the Brain-like Computing and Applications,
School of Information Science and Engineering, Xiamen University, Xiamen 361005, China)

Abstract: Because of the complexity of the author's attribution, it is difficult to use the traditional natural language processing model to complete the authorship identification. To discover the author's attribution, we use the deep model of the denoising autoencoder to analyze the text structure and identify the author's writing style in the text, and the SVM classifier is used to accomplish the recognition of authors. The advantage of the model lies in considering the noise diversity and complexity of unknown text features, and it can reconstruct the original text input with noise. This method is applied to the recognition of poetry authors such as Wu Chengen, Wang Tingchen, Xue Hui, etc. The most accuracy of recognition is 78.2%, it verifies the validity of the method. Furthermore this method is applied to the identification of poetry authors in "Journey to the West".

Key words: denoising autoencoder; code feature; authorship identification