适用于特殊类型自然语言分类的自适应特征谱神经网络

王一峰，孙丽茹，崔良乐，赵毅
哈尔滨工业大学（深圳）理学院，广东 深圳 518055

摘要
计算机算力的提升使得深度学习算法迅速发展，然而由于古诗文特殊的语序、用词、结构、句式、文法结构、表达方式，深度学习模型需要消耗更多的算力进行特征提取等工作，因此并未在这一领域取得广泛的应用。为此，提出了一种新型的神经网络结构——自适应特征谱神经网络。该算法有效减少了运算时间，可以自适应地选择对分类最有用的特征，形成最高效的特征谱，得到的分类结果具有一定的可解释性，而且由于其运行速度快、内存占用小，因此非常适用于学习辅助软件等方面。以此算法为基础，开发了相应的个性化学习平台。该算法使古诗文分类的准确率由93.84%提升到了99%。

关键词
自适应特征谱；神经网络；文本分类；古诗词；拉普拉斯矩阵

中图分类号：TP183，O29，TP312 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020036

Adaptive feature spectrum neural networks for special types of natural language classification

WANG Yifeng, SUN Liru, CUI Liangle, ZHAO Yi
School of Science, Harbin Institute of Technology(Shenzhen), Shenzhen 518055, China

Abstract
The improvement of computer computing power has led to the rapid development of deep learning algorithms. However, due to the special word order, wording, structure, sentence structure, grammatical structure, and expression of ancient poetry, deep learning models need to consume more computing power for feature extraction, etc. Therefore, it has not been widely used in this field. As a result, a new kind neural network: the adaptive feature spectrum neural network was proposed, which can considerably reduce the computation and adaptively select the features that are the most useful for classification in order to form the most efficient feature spectrum. The classification results obtained have certain interpretability. Moreover, its fast running speed and lower RAM consumption make it very suitable for learning aids software, and other fields. Based on this algorithm, a corresponding personalized learning platform was developed. This algorithm improves the classification accuracy of ancient Chinese poetry from 93.84% to 99%.

Key words
adaptive feature spectrum, neural network, text classification, ancient poems, laplace matrix

1 引言

文本分类问题是自然语言处理领域一个十分常见的问题，文本分类应用非常广泛，例如舆情分析、影评分析、新闻情感分析[1-2]、新闻内容分类[3-4]、垃圾邮件过滤、敏感信息自动屏蔽、社交软件交流中对某句话的情感趋势分析，以及购物网站中的“好评度”评估。总而言之，语言本身是一种人类智慧的体现，而文本作为语言的载体，蕴含着大量的信息和规律，因此让计算机掌握这种规律并进行模式识别和分类是一项对算法的巨大挑战。而古诗文作为一种特殊的语言形式，其表达方式与现代语言相比更加隐晦、精练，与白话文相比分类难度更高，因此本文选择古诗文分类问题作为文本分类的切入点，以便提出更优的文本分类算法。

文本分类算法是自然语言处理中很重要的一类算法，在20世纪50年代就已经有科学家借助“专家系统”对文本进行分类[5]，然而该方法可覆盖的范围以及分类准确率都非常有限，只能用于解决一些条件明确、描述清晰且有条理的文本分类问题。随着统计学方法的发展，特别是20世纪90年代后互联网在线文本数量的增长和机器学习学科的兴起，逐渐形成了一套解决大规模文本分类问题的经典方法[6]，其主要流程是“人工特征工程+分类器”，即把整个文本分类问题拆分成特征工程和分类器两部分。对于不同类型的文本，特征选取方法是不同的，分类器的设计也是不同的，例如：采用Apriori算法对同时出现在语句中的特征项进行筛选，进而实现分类[7]；基于遗传算法对诗文特征项进行选取，接着利用朴素贝叶斯模型进行分类[8]；通过均值漂移、谱聚类、k-means等聚类算法选取特征，随后采用支持向量机[9]、距离加权最近邻、贝叶斯模型等分类器进行分类[10-11]。其中，使用聚类算法寻找特征，随后采用加权最近邻分类器的方法是目前对中国古诗文分类准确率最高的一种方法，平均准确率可以达到93.84%[12]，其中，针对某一特定类型古诗词文本的分类准确率最高可以达到96.67%。

然而这些分类方法存在几个主要缺点。首先，现有的古诗文本分类算法的性能依赖于初始特征库的选取，以专家选取的特征库为基础进行特征聚类、文本分类的性能远好于以普通人选取的特征库为基础的性能。除此之外，找特征的过程与分类的过程往往是分离的，这会导致一些被选取的特征对分类任务作用不大[13-15]，应考虑将古诗文分类的结果直接反馈到找特征的过程，进而帮助找到更好的分类特征。这些缺陷最终导致在面对不同类型的文本，尤其是面对语言委婉、内容写意、抒情的文本时，难以设计出效果良好的分类器。因此本文设计了自适应特征谱神经网络来完成文本分类任务，它可以自适应地选择对分类有效的特征，并组成“最优特征谱”。

2 数据预处理

在机器学习算法中，输入的数据通常是数值型的，因此需要将文字型文本特征转换为数值型数据特征，将输入模型的文本变成向量，从而确保模型可以进行计算和分类。

具体操作是用高频词组成特征库，再将特征库中的词用向量表示。首先，使用Sunday算法[16]查找古诗文中出现频率较高的字词，组成“特征库”。使用Sunday算法的好处是在字符串匹配时可以大幅减少运算时间。Sunday算法查找原理如**图1**所示。

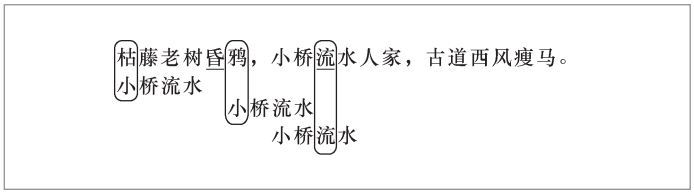


图1 Sunday算法查找原理

任务是判断文本串“枯藤老树昏鸦，小桥流水人家，古道西风瘦马。”中是否包含模式串“小桥流水”。从左端开始，比较模式串的第一位，如果匹配，则依次向下比较；若不匹配，则比较文本串中下一字符是否出现在模式串中（本例中“昏”并未出现在模式串“小桥流水”中），因此直接向右平移m+1个单位（m为模式串长度）。重复上述过程，若发现文本串的下一字符出现在模式串中：本例中“流”出现在“小桥流水”中，这时直接将两个“流”对齐，再逐位比较，最终发现匹配成功。

由于古诗文中单音节词占多数，且文法注重典故、骈骊对仗、音律工整，因此在内容表达上就会有一些牺牲。此外，一些在现代文中并不多见的特征词（如“金樽”“涧户”“左迁”等）在古诗文中却并不罕见，现代文的分词方法有时很难将其准确分开，因此在借助Sunday算法进行词频统计的基础上，还需要进行一些人工的筛选，这也是本文的一项重要工作。

在得到由高频字词组成的特征库后，要进行更精细化的筛选。目标是将输入的古诗文分成4类，因此特征词的选择标准应与该特征词对4种类型古诗文本的区分表示度相关。有些字词虽然出现频率高，但对于分类而言用处不大。按照爱情、忧国忧民、山水田园、哲理诗的顺序，从4类诗中各选取一句话：“愿得一心人，白头不相离”“秦时明月汉时关，万里长征人未还”“涧户寂无人，纷纷开且落”“人生得意须尽欢，莫使金樽空对月”。若直接将文本的出现频次作为文本分类的特征输入，会发现4类诗中均出现了“人”字，而“月”字则出现了两次。“人”和“月”看起来似乎是很重要的两个特征，但事实上，这两个词是比较常见的、不具备区分能力的词，很多诗篇会用到，因此不能单纯地选取文本的词频来反映诗的特征，而诸如“白头”“长征”“涧户”“金樽”等仅出现一次的词反而更能反映其类别特征。因此，使用词频-逆文本频率（term frequency-inverse document frequency，TF-IDF）方法对其进行向量表示[17]。

设爱情类、山水田园类、忧国忧民类、哲理类古诗分别对应类别1、类别2、类别3、类别4，每种类别下对应的篇数分别为N_1、N_2、N_3、N_4，第i类下第j篇古诗文包含的汉字总数目为$n_{i,j}$，特征词t在该篇诗文中出现次数为$n_{t,i,j}$（i=1，2，3，4，j=1，2，…，N_i），则特征词t在第i类文本中的词频$\mathrm{TF}_{t,i}$为：

$$\mathrm{TF}_{t,i}=\frac{\sum_{j=1}^{N_i} n_{t,i,j}}{\sum_{j=1}^{N_i} n_{i,j}} \tag{1}$$

$\mathrm{TF}_{t,i}$表示特征词t在第i类文本中的出现率，同时也是对词数的归一化，以避免其偏向更长的文本文件。

逆文本频率（IDF）是对某个特征词的“普遍重要性”的度量。设所有文本中包含特征词t的篇数为DF_t，所有文本数量为$N=N_1+N_2+N_3+N_4$，则特征词t的IDF_t为：

$$\mathrm{IDF}_t=\log\frac{N}{\mathrm{DF}_t} \tag{2}$$

因此，特征项t的$\mathbf{TF_IDF}_t$表示一个1×4的向量：

$$\mathbf{TF\_IDF}_t=(\mathrm{TF}_{t,1},\ \mathrm{TF}_{t,2},\ \mathrm{TF}_{t,3},\ \mathrm{TF}_{t,4})\times\mathrm{IDF}_t \tag{3}$$

筛选标准是向量$\mathbf{TF_IDF}_t$的标准差

$\sigma(\mathbf{TF\_IDF}_t)$：

$$E(\mathbf{TF\_IDF}_t)=\frac{1}{4}\sum_{i=1}^{4}\mathrm{TF}_{t,i}\times\mathrm{IDF}_t \quad (4)$$

$$\sigma(\mathbf{TF\_IDF}_t)=\sqrt{\frac{1}{4}\sum_{i=1}^{4}(\mathrm{TF}_{t,i}\times\mathrm{IDF}_t-E(\mathbf{TF\_IDF}_t))^2} \quad (5)$$

$\sigma(\mathbf{TF_IDF}_t)$较大的特征词对特定类型的古诗文有更强的表示能力。该做法的主要思想是：如果一个词在某一类文本中出现频率很高，而在所有文本中出现频率却不高，那么该词对于这类文本就具有很强的代表性和区分度[18]，反之亦然。因此可以过滤一些常见的词语，保留重要的词语，从而实现特征词的精细化提取。

下一步需要将最终筛选出的特征词进行向量化表示。现有的古诗文本分类研究多采用TF-IDF方法进行特征词的向量化表示，并且取得了90%以上的准确率。词嵌入（word embedding）表示被提出后，文本分类问题逐渐向基于词嵌入表示或词向量的方法展开研究，如之前基于卷积神经网络（convolutional neural network，CNN）的文本分类方法[19]以及近期基于Transformer的文本分类方法[20]。本文对以下两类方法进行了融合，TF-IDF表示方法具有更强的可解释性，并且在古诗文分类领域使用时间较长，而词嵌入表示方法则在近年来被广泛应用于自然语言处理领域，借助深度学习模型强大的性能，其表示效果得到了广泛的认可。

借助古文、白话文识别任务来完成特征词嵌入表示。与古诗词主题分类不同，古文、白话文识别任务的数据集更加方便易得，且标签也更易标注。采用连续词袋（continuous bag-of-words，CBOW）模型[21]将特征词转化为1×100的向量，并取其中的5个维度进行可视化，如**图2**所示。

从**图2**可以看到，位置相近、大小相近、颜色相近的特征词具有更加相近的含义。设由CBOW模型得到的特征词t的词向量为$\mathbf{vector}_{t_\mathrm{CBOW}}$，则特征词$t$的最终表示向量$\mathbf{vector}_t$为：

$$\mathbf{vector}_t=\mathbf{vector}_{t\_\mathrm{CBOW}}\times\sigma(\mathbf{TF\_IDF}_t) \quad (6)$$

其中，$\sigma(\mathbf{TF_IDF}_t)$为$\mathbf{TF_IDF}_t$向量的标准差。最终得到的词向量$\mathbf{vector}_t$不仅包含特征词的语义信息，同时也包含该特征词对分类任务的重要度评价，在自然语言处理领域的很多研究中[22]，有将词频-逆文本频率信息作为权重进而构造词典的范例。因此将结合了TF-IDF方法与CBOW方法得到的词向量$\mathbf{vector}_t$作为最终的特征词表示结果。

3 自适应特征谱神经网络的构造

由于古文的句式、格式、表达方式都有别于现代文，且单音节词占多数，一篇古文包含的特征词数量繁多，如果使用传统的神经网络模型进行分类，计算规模将非常庞大。为了使算法可以更方便地搭载于手机、学习机等终端设备之上，进而使得基于该算法的软件成为广泛的学习平台，除了分类准确率之外，对内存占用、运行速度也有一定的要求[23]。同时，为了满足教育大数据、辅助学习软件的需要，应在一定程度上对最终的分类结果进行解释，或者对特征选择进行一定程度的可视化。因此，笔者设计了自适应特征谱神经网络，它可以对众多特征词进行筛选，自适应地形成对分类最有意义的特征谱，而后只需在输入的古诗文中进行检索，将特征谱中对应的特征词提取出来，并乘以对应权重，然后将结果输入后续神经网络，即可得到分类结果。

特征是对数据内在规律的反映，而对特征之间相互关系的理解与升华则是文本

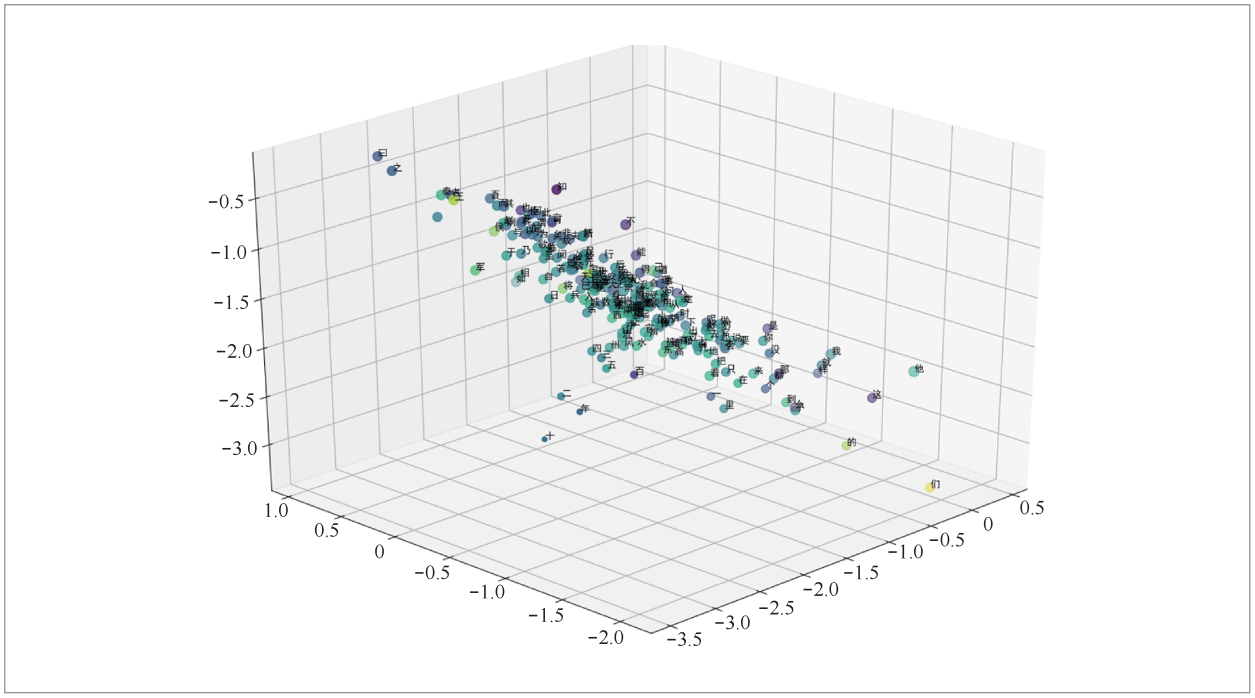


图 2 基于 CBOW 模型的古诗文特征词向量可视化

大数据语义理解的重要手段[24]。本文提出的自适应特征谱神经网络将特征词之间的相互关系融合在拉普拉斯矩阵中。拉普拉斯矩阵是一种图的矩阵表示形式，描述了图中各节点之间的关系。文本分类任务一般是通过对不同特征的相互耦合来完成的，因此，对特征与特征之间关系的描述正是其所需要的。下面将拉普拉斯矩阵的一部分作为神经网络的输入层。

为了得到拉普拉斯矩阵$\boldsymbol{L}$，首先需要计算各特征项的相似度矩阵$\boldsymbol{A}$，其中$A_{i,j}=\cos(t_i,t_j)$，这里采用余弦相似度来表征特征项t_i和特征项t_j的相似度。进而可以构建对角矩阵$\boldsymbol{D}$，其中对角元素D_{ii}为：

$$D_{ii}=\sum_{j=1}^{n}A_{i,j} \tag{7}$$

则拉普拉斯矩阵$\boldsymbol{L}$可表示为：

$$\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{8}$$

由于拉普拉斯矩阵是对称矩阵，因此只需将其上三角部分取出作为拉普拉斯层即可。

接下来构造自适应特征谱神网络的核心结构——自适应特征谱层。它由拉普拉斯层经过全连接网络（全连接网络就是层与层之间的计算过程，即把前一层与后一层的节点全部相连）得到。整体的网络结构如图3所示。

图3展示了自适应特征谱神经网络的训练过程。首先，拉普拉斯层记载着特征项之间的全部关系，后接一个全连接网络，旨在输出最优的特征谱，后续的神经网络结构将以该特征谱为基础完成文本分类任务。设特征库中有n个特征项，这里设定在特征谱中只保留m个特征项（$m<n$），使得神经网络留下对分类最有用的特征。如果前期负责生成特征谱的网络工作效果不佳，将导致后续文本分类效果不佳，因此对误差函数做反向传播（back propagation，BP），既调整了分类网络，也

调整了特征生成网络。这就解决了前文提到的分类器与特征选择工作分离而导致效率不高的问题，因此称之为“自适应特征谱”。为了缓解训练过程中的过拟合问题，在该全连接网络中进行了Dropout操作，以减少特征检测器（神经元节点）间的相互作用，达到正则化的效果，本文将Dropout比率设置为0.5。

此外，特征谱层还减小了特征数量，降低了对算力的损耗。因此本文提出的自适应特征谱神经网络算法适合处理复杂的文本分类问题，即使输入海量的数据，运算量也不会过大，这是因为要求特征谱层只能保留一定量的、对分类最有用的特征，对分类最有用的特征并非像传统方法那样由人为因素决定，而是完全通过大量数据自主训练得到的。自适应特征谱神经网络算法的分类准确率会随着输入特征的增多而提高。

完成神经网络的训练后，得到了现阶段对分类最有意义的特征谱，被称为“最优特征谱”。由于拉普拉斯层与自适应特征谱层之间的网络结构已经完成了根据分类任务筛选特征、给出相应权重的任务，因此在测试或应用时，只保留最优特征谱及后续的输入层、隐藏层、输出层结构，这大大缩短了实际应用时的响应时间。以最优特征谱为基础，对每篇古诗文对应的表示向量做如下操作：用Sunday算法在输入文本中搜索最终保留的m个特征词，假设检索到了k个特征词（$k \leqslant m$），则对这k个特征词对应的特征谱中的数值进行归一化，之后分别乘以这k个特征词的词向量，最终再对这k个词向量求和。这种方法的本质是以k个特征词在最优特征谱中对应的数值为基础，对其对应的词向量进行加权平均，最终得到可以表示输入文本的文本向量。借助这种方法，该模型的输入维度始终可以保持为词向量的维度，运行速度、内存占用并不会随着输入文本长度的增加而发生明显变化。

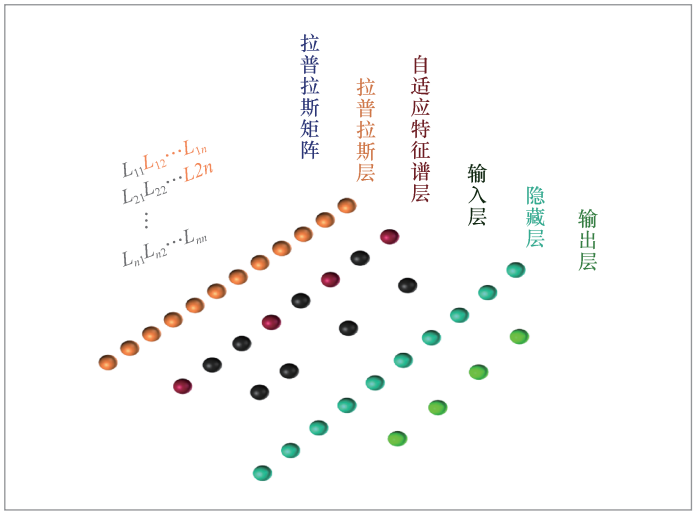


图 3 自适应特征谱神经网络结构

4 实验结果与改进空间

本文所用的古诗文数据均来自“古诗文网”，该网站将所有古诗都进行了分类，本文将该网站对古诗的分类类别作为每首古诗对应的标签，并输入本文设计的自适应特征谱神经网络进行训练，得到最终的结果。

为了验证本文提出的自适应特征谱神经网络算法的准确率，进行了两次对比实验，使用的是目前对古诗文本分类准确率很高的两种方法：基于谱聚类算法的特征聚类+加权最近邻分类器；基于k-means算法的特征聚类+加权最近邻分类器。两种方法都以预先选定的特征库为基础，对其中的特征项进行聚类分析。其出发点在于每个特征对每一类型文本的表示能力不同，例如：出现“鸳”字的文本有较大概率是以爱情为主题的；“田”“园”等字则对山水田园类文本区分度较高；“烛”字对爱情类、哲理类文本都有不错的表示度。将不同特征词对不同类型文本的表示能力可视化，爱情类、山水田园类文本的表示能力可视化分别如**图4**、**图5**所示。

图 4 不同特征词对爱情类文本的表示能力可视化

图 5 不同特征词对山水田园类文本的表示能力可视化

图6表示同时将400个特征词对4个维度的信息进行了可视化：爱情类为竖轴，山水田园类为横轴，在坐标轴上的投影值越大表示对该类型文本的表示能力越强；圆圈的大小表示该特征词对忧国忧民类古诗文本的表示能力，颜色深浅表示该特征词对哲理类古诗文本的表示能力。

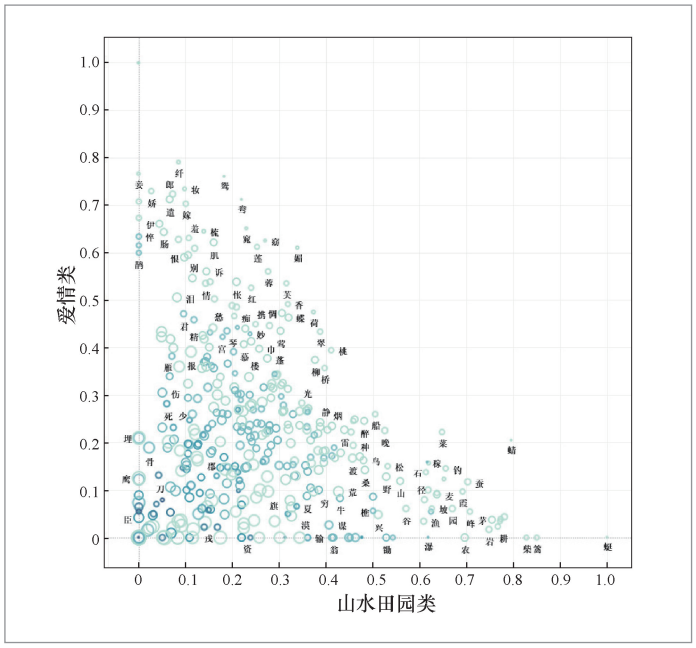


图 6 400 个特征词对 4 个类型文本的表示能力可视化总览

两种聚类算法都是根据每个特征项对不同类别文本表示能力的强弱，将特征项分为4类，再将聚类结果交给后续的分类器，使用加权最近邻算法完成对文本的分类。

不同聚类算法的聚类结果是不同的，**图7**、**图8**分别为k-means算法聚类结果中的忧国忧民类、爱情类文本特征，**图9**、**图10**分别为谱聚类算法聚类结果中的哲理类、山水田园类文本特征。

从聚类算法得到的结果可以看出，不同类别的特征词基本是符合常识的，但仍然有许多字是“模棱两可”的。而且聚类一旦完成，其结果无法根据后续分类器的表现随时进行修改和调整，这是此类算法的弊病。而本文提出的自适应特征谱神经网络则能够消除类似的“断层”现象，特征谱结构随时服务于分类器结构，并在训练的过程中不断更新，结果自然更加理想。

将1 500篇古诗文作为训练集，将500篇古诗文作为测试集（每类125篇），分别使用谱聚类算法+加权最近邻分类器、k-means算法+加权最近邻分类器、普通的含有两层隐藏层的反向传播神经网络和本文提出的自适应特征谱神经网络4种算法对4类古诗文本进行分类测试，得到的结果见**表1**。

哲理类诗文相比其他类型更难分类，而且它们覆盖的范围本身也有交叠。例如，很多田园归隐诗本身就有看破世事、论述自己的人生思考、抒发哲理的部分在内，因此哲理类古诗文的分类准确率一般低于其他3种类型。

从**表1**可以看到，BP神经网络对古诗文本的分类准确率明显高于各种聚类算法+传统分类器的组合，而本文提出的自适应

特征谱神经网络可以使分类准确率上升到99%，在某些特定类别上甚至可以达到不出错的程度（当然，这和本文测试集数量太少有关，这也是未来改进的方向）。

5 性能分析及应用

本文提出的自适应特征谱神经网络的性能优势在于它可以自适应地选择最有助于分类任务的特征词。通过特征词向量构建的拉普拉斯层记录不同特征词之间的相互关系，而后边的全连接网络则是对这种特征词之间相互关系的整合。该网络结构对不同输入文本的文字组合，赋予的特征词权重各不相同，特征词将以该权重为基础，参与下一阶段的运算，最终得到该文本的主题分类结果。当分类错误时，误差会通过整体的网络结构进行反向传播，并追溯到此前赋予特征词的权重，而这些特征词的权重以及网络结构中的其他参数则会通过梯度下降算法进行更新，并参与下一个循环的计算。神经网络模型就是以此来完成对特征词权重的学习的，该学习过程是一种“自适应”的调整过程。

在完成大量的迭代计算之后，自适应特征谱神经网络得到了充分的训练。训练完成的自适应特征谱神经网络会对不同的输入文本提取不同的特征词，并为其分配不同的权重。以古诗文《孔雀东南飞》为例，自适应特征谱神经网络根据不同文字的组合方式，对文中有助于主题分类的特征词进行提取，并为其分配了适当的权重，该权重经过后续网络结构的运算即可得到最终的分类结果。按照文本中不同特征词权重的数值，生成《孔雀东南飞》的专属特征词词云图，如图11所示。特征词在词云图中的大小与其被自适应特征谱神经网络赋

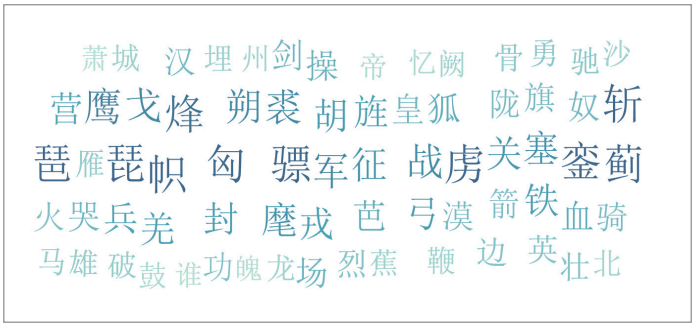


图 7 k–means 算法聚类结果：忧国忧民类文本特征

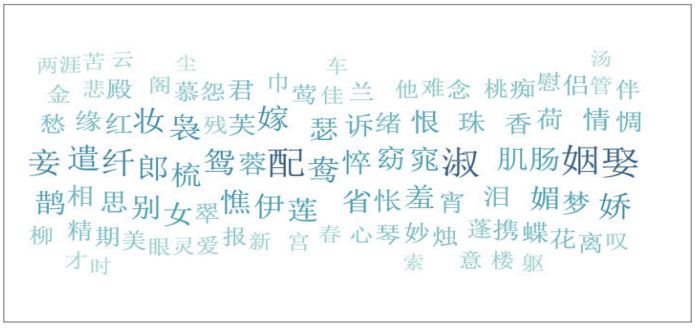


图 8 k–means 算法聚类结果：爱情类文本特征

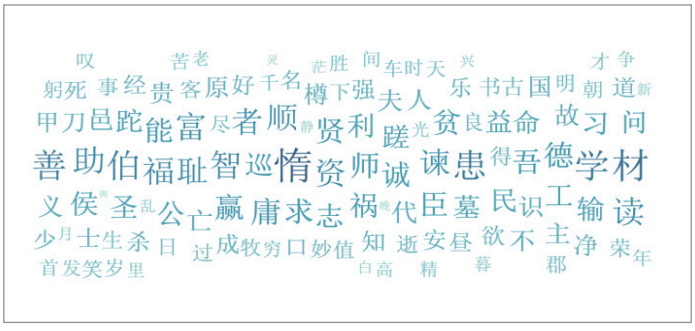


图 9 谱聚类算法聚类结果：哲理类文本特征

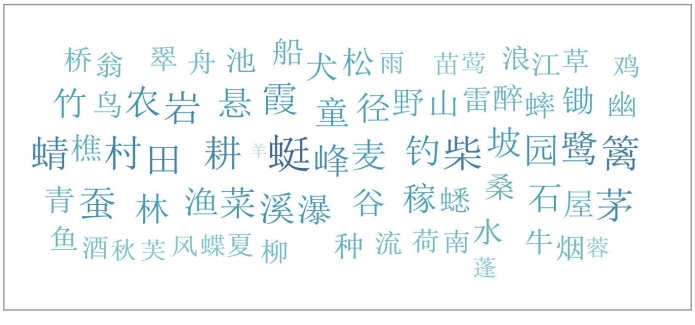


图 10 谱聚类算法聚类结果：山水田园类文本特征

表1 4种分类算法的分类准确率对比

模型	爱情诗	山水田园诗	哲理诗	忧国忧民诗	平均
谱聚类算法+加权最近邻分类器	78.4%	88.0%	69.6%	62.4%	74.6%
k-means聚类算法+加权最近邻分类器	85.6%	94.4%	52.6%	59.2%	72.9%
BP神经网络	98.8%	98.2%	96.0%	99.2%	98.3%
自适应特征谱神经网络	99.8%	99.4%	98.4%	99.6%	99.3%

予的权重成正比。

从图11可以看出，自适应特征谱神经网络为《孔雀东南飞》中的“姻”“缘”“情”“相”“望”等字赋予了较大的权重，这些特征词与“爱情”的联系确实较为密切，符合人们的日常认知。然而同样和“爱情”关系密切的“妻”“女”“君”“愁”“怨”“泪”“嫁”等字则未被赋予太大的权重，这说明自适应特征谱神经网络认为这些字在文中的出现形式和组合方式对“爱情”这一主题的代表性不强，或者这些字的出现方式与家国类、哲理类等其他类型的古诗文类似，因此只被赋予了中等大小的权重。而诸如“安”“山”“水”等字，自适应特征谱神经网络认为其与“爱情”并无明显关系，因此对其赋予的权重极小，这也与人们的日常认知相符。由此可见，自适应特征谱神经网络确实可以自适应地提取有助于分类的特征。

图11 《孔雀东南飞》特征词词云图

自适应地提取特征可以避免在分类时局限于固定的特征词，它可以根据文本中不同文字之间的组合关系进行适当的调整，即使对于同一类型的古诗文，特征词的选取也有很大的分别。本文在爱情类、哲理类、山水田园类3类古诗文中各选取了4篇，并根据自适应特征谱神经网络给出的特征词权重分别绘制了词云图，如图12所示。这12个词云图分别由3个类别的12首不同的古诗词生成，从这12个词云图可以看出，对于不同的古诗词文本来说，自适应特征谱神经网络关注的特征词也并不相同，即使对于同一类型的古诗文，自适应特征谱神经网络给出的特征词权重也各不相同。除此之外，由于自适应特征谱神经网络提取特征的能力是在数据集中训练获得的，因此随着数据量和算力的增长，自适应特征谱神经网络对特征的提取能力以及对文本的分类能力也会显著提升。

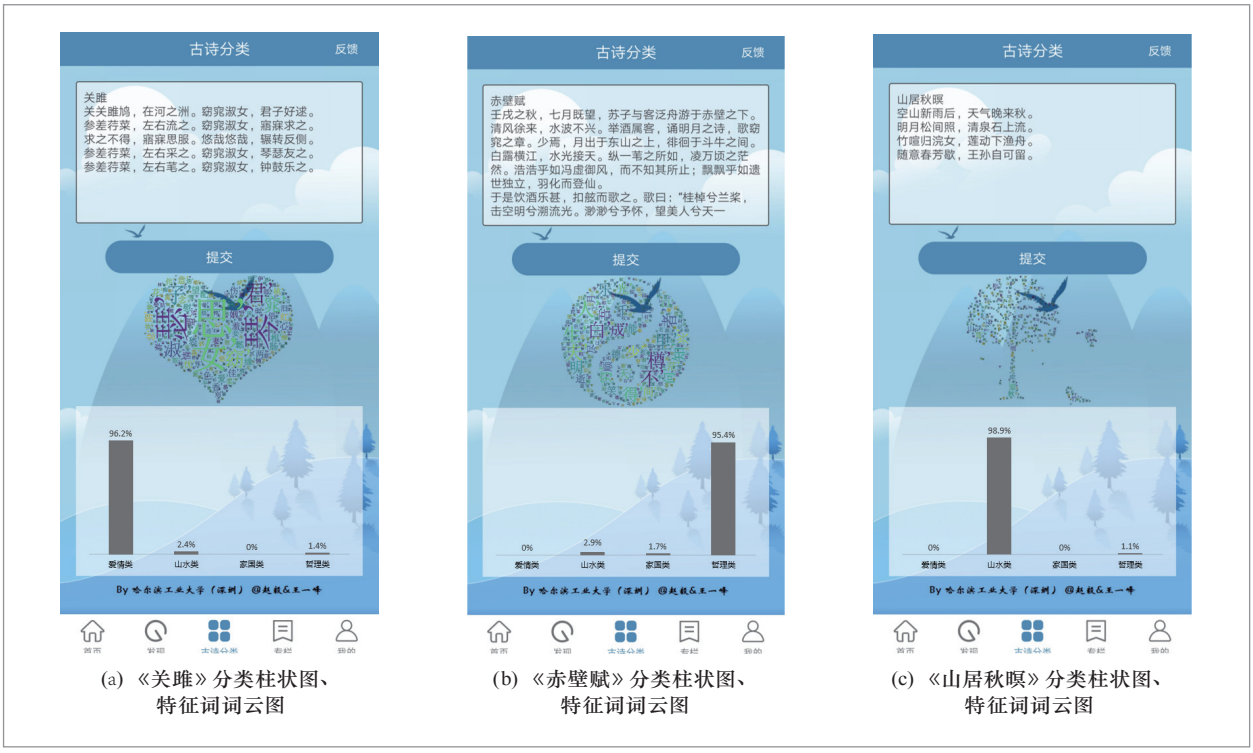
目前，笔者已经以自适应特征谱神经网络算法为基础，与相关科技公司合作，开发了一款古诗文主题分类App，如图13所示。当用户在文本框中输入古诗文内容后，下方会显示该古诗文所属不同类别的概率柱状图，同时会输出一张由自适应特征谱神经网络给出的特征词权重所生成的词云图作为分类依据。由此可见，该算法具有一定的创新性，并且在市场上具有广阔的应用前景。

图 12 自适应特征谱神经网络对不同古诗文本提取特征词所生成的词云图

6 结束语

本文提出的自适应特征谱神经网络的设计灵感来源于谱聚类算法，然而在完成网络结构的设计之后，笔者发现其结构和卷积神经网络有些相似之处，例如，用卷积层、池化层处理图像数据的初衷是将输入的图像数据降维，并提取合适的特征，该特征并非人工提取，而是根据所要完成的任务以及误差情况自动提取的；而自适应特征谱层也是为了将输入的表示向量降维，删除其中不重要的特征，选取合适的特征，这种选取不受人为干预，而是将训练过程中产生的误差进行反向传播，自适应地进行调整。卷积神经网络非常适用于处理图像类型的数据，而本文提出的自适应特征谱神经网络则非常适用于处理文本数据。因此，本文提出的网络结构具有非常广阔的应用前景。

此外，使用本文提出的自适应特征谱神经网络进行特征提取以及分类得到的结果具有一定的可解释性，且在实际应用时响应速度快、内存占用小，因此非常适合用于辅助教育平台的开发，基于该算法开发的古诗文主题分类App受到了用户的一致好评。用户在使用该App时，无疑也提供了海量的训练样本，以此为基础，笔者可以继续优化该模型，达到更高的分类精度。以“更大的数据”驱动“更好的深度学习模型”正是后期优化的方向。



(a) 《关雎》分类柱状图、特征词词云图　(b) 《赤壁赋》分类柱状图、特征词词云图　(c) 《山居秋暝》分类柱状图、特征词词云图

图 13　基于自适应特征谱神经网络开发的古诗文主题分类 App

参考文献:

[1] 李钝, 曹付元, 曹元大, 等. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008, 35(4): 132–134.
LI D, CAO F Y, CAO Y D, et al. Text sentiment classification based on phrase patterns[J]. Computer Science, 2008, 35(4): 132–134.

[2] 胡熠, 陆汝占, 李学宁, 等. 基于语言建模的文本情感分类研究[J]. 计算机研究与发展, 2007, 44(9): 1469–1475.
HU Y, LU R Z, LI X N, et al. Research on language modeling based sentiment classification of text[J]. Journal of Computer Research and Development, 2007, 44(9): 1469–1475.

[3] 沈加. 基于SVM模型的新闻分类系统设计与实现[D]. 成都: 电子科技大学, 2013.
SHEN J. The design and realization of webnews classification system based on SVM[D]. Chengdu: University of Electronic Science and Technology of China, 2013.

[4] 潘澄. 基于领域向量模型的新闻网页分类算法[J]. 软件导刊, 2015(7): 57–60.
PAN C. News web classification algorithm based on domain vector model[J]. Software Guide, 2015(7): 57–60.

[5] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1–47.

[6] 郝春风, 王忠民. 一种用于大规模文本分类的特征表示方法[J]. 计算机工程与应用, 2006, 43(15): 170–172.
HAO C F, WANG Z M. Method of expressing features used for large-scale text classification[J]. Computer Engineering and Applications, 2006, 43(15): 170–172.

[7] 吴春龙, 周昌乐. 基于频繁关键字共现的诗词风格分类模型研究[J]. 厦门大学学报(自然科学版), 2008, 47(1): 41–44.
WU C L, ZHOU C L. Frequent keyword concurrence-based vector space model

for Chinese poetry style analysis[J]. Journal of Xiamen University (Natural Science), 2008, 47(1): 41–44.

[8] 孙晋文, 肖建国. 基于SVM的中文文本分类反馈学习技术的研究[J]. 控制与决策, 2004, 19(8): 927–930.
SUN J W, XIAO J G. Study on feedback learning of SVM-based Chinese text classification[J]. Control and Decision, 2004, 19(8): 927–930.

[9] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features[C]//European Conference on Machine Learning. Heidelberg: Springer, 1999: 137–142.

[10] KIM H, HOWLAND P, PARK H, et al. Dimension reduction in text classification with support vector machines[J]. Journal of Machine Learning Research, 2005, 6(1): 37–53.

[11] 易勇, 何中市, 李良炎, 等. 基于遗传算法改进诗词风格判别的研究[J]. 计算机科学, 2005, 32(7): 156–158.
YI Y, HE Z S, LI L Y, et al. A traditional Chinese poetry style identification calculation improvement model[J]. Computer Science, 2005, 32(7): 156–158.

[12] 黄永锋, 李奇. 基于特征项聚合的古典诗歌分类模型[J]. 东华大学学报(自然科学版), 2014, 40(5): 599–604.
HUANG Y F, LI Q. Classical poetry classification model based on feature terms clustered[J]. Journal of Donghua University (Natural Science Edition), 2014, 40(5): 599–604.

[13] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 18–24.
ZHOU Q, ZHAO M S, HU M. Study on feature selection in Chinese text categorization[J]. Journal of Chinese Information Processing, 2004, 18(3): 18–24.

[14] 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.
DAI L L, HUANG H Y, CHEN Z X. A comparative study on feature selection in Chinese text categorization[J]. Journal of Chinese Information Processing, 2004, 18(1): 26–32.

[15] 单丽莉, 刘秉权, 孙承杰. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报, 2011(S1): 319–324.
SHAN L L, LIU B Q, SUN C J. Comparison and improvement of feature selection method for text categorization[J]. Journal of Harbin Institute of Technology, 2011(S1): 319–324.

[16] 潘冠桦, 张兴忠. Sunday算法效率分析[J]. 计算机应用, 2012, 32(11): 3082–3088.
PAN G H, ZHANG X Z. Study on efficiency of Sunday algorithm[J]. Journal of Computer Applications, 2012, 32(11): 3082–3088.

[17] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J]. 计算机学报, 2011(5): 856–864.
HUANG C H, YIN J, HOU F. A text similarity measurement combining word semantic information with TF-IDF method[J]. Chinese Journal of Computers, 2011(5): 856–864.

[18] 范珈瑜. 基于文本挖掘的游客对古镇旅游态度的分析[J]. 大数据, 2017, 3(6): 93–101.
FAN J Y. Analysis of tourists' attitude for ancient towns based on text mining[J]. Big Data Research, 2017, 3(6): 93–101.

[19] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]//The 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 649–657.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. Computer Science, 2018, arXiv: 1810.04805.

[21] 程学旗, 兰艳艳. 网络大数据的文本内容分析[J]. 大数据, 2015, 1(3): 62–71.
CHENG X Q, LAN Y Y. Text content analysis for web big data[J]. Big Data Research, 2015, 1(3): 62–71.

[22] 宋云生. 一种情感判别分析体系在汽车品牌舆情管理中的应用[J]. 大数据, 2017, 3(6): 55–64.
SONG Y S. Application of an emotion

discriminant analysis system in the management of automobile brand[J]. Big Data Research, 2017, 3(6): 55–64.

[23] 吴毅坚, 陈士壮, 葛佳丽, 等. 数据自治开放的软件开发和运行环境[J]. 大数据, 2018, 4(2): 31–41.
WU Y J, CHEN S Z, GE J L, et al. Software development and runtime environment for self-governing openness of data[J]. Big Data Research, 2018, 4(2): 31–41.

[24] 袁书寒, 向阳, 鄂世嘉. 基于特征学习的文本大数据内容理解及其发展趋势[J]. 大数据, 2015, 1(3): 72–81.
YUAN S H, XIANG Y, E S J. Text big data content understanding and development trend based on feature learning[J]. Big Data Research, 2015, 1(3): 72–81.

作者简介

王一峰（1995– ），男，哈尔滨工业大学（深圳）理学院硕士生，主要研究方向为自然语言处理、计算机视觉、智能控制、机器人运动、惯性制导以及机器学习的数学原理。

孙丽茹（1994– ），女，哈尔滨工业大学（深圳）理学院硕士生，主要研究方向为自然语言处理、教育大数据和机器学习中的聚类算法。

崔良乐（1978– ），男，哈尔滨工业大学（深圳）理学院讲师，主要研究方向为西方美学、中国近现代思想文化传播、文化研究和与在线学习相关的教育大数据。

赵毅（1977– ），男，博士，哈尔滨工业大学（深圳）理学院教授、博士生导师，哈尔滨工业大学（深圳）应用数学研究中心主任，主要研究方向为非线性时间序列分析、动力系统、复杂网络、生物数学和数据科学。

收稿日期： 2019-12-10

基金项目： 学位与研究生教育资助项目（No. 2017Y0902）；深圳市教育科学规划2015年度重大招标课题重点资助项目（No. zdzz15001）；哈尔滨工业大学（深圳）高等教育教学改革资助项目

Foundation Items: Academic Degrees & Graduate Education Program (No. 2017Y0902), Shenzhen Education Science Planning 2015 Major Bidding Project Key Funded Project (No. zdzz15001), Harbin Institute of Technology (Shenzhen) Higher Education Teaching Reform Project