基于神经网络算法对文本的情感分析

王钟璞

（河南省开封高级中学，河南开封，475000）

摘要：由于互联网的出现，网络上有了大量的文本信息，例如新闻、社交软件的聊天记录以及各种评论，但是有文字就少不了情感分析，因为人脑分析的不准确，并且会夹杂了一些主观的情绪，所以我们选择用机器代替人进行分析，但是计算机不是生来就会情感分析，所以在编程的时候需要一些心理学的知识和生活中积累的经验。只有这样计算机才能够理解并学会处理各种不同情绪，进而帮助人类解决繁杂的事务性工作。本文通过利用发展迅速的神经网络模型，将神经网络模型应用到文本的情感分析上，无疑是一种领先的尝试，经过试验验证得知，模型的准确率达到了96%以上的准确率。

关键词：神经网络算法；情感分析；词向量；文本分析

DOI:10.16589/j.cnki.cn11-3571/tn.2020.20.014

0 引言

据有关数据统计，仅在中国，每天会有百亿条信息被发出[1]，这些信息如果只让人工去分析，工作量会很巨大。如果使用电子计算机来完成这项工作，会省去很多时间，但电子计算机来代替人类进行情感分析，也有一些弊端。因为电子计算机本身是没有情感的，只能人们去给它编辑情感，因为人的情感是多变的，机器只能局限于所编辑的情感。

现在的人工智能也只能做一些简单的情感分析，如果这个作品可以更好地改善，即可以运用在很多人工智能作品上，以后人工智能就会有人的情感（除了自私、造反等负面情感），更好地为人类服务。

神经网络是一门发展相对长远的学科[2]。上世纪 80 年代就已经兴起了，然而在短短十年内衰落了，其原因是由于计算量偏大的计算方式并不适用于当时的计算机。可随着人工智能、大数据时代的到来，各种超级计算机相继面世，这一难题于是得以解决。由于是模拟人脑，神经网络拥有很多优秀的特性，例如：数据处理能力强，能够针对大量数据进行处理，且数据越多，表现越好；计算能力强，相比于人脑，神经网络能够在单位时间内爆发出指数级的计算能力；学习程度高，更能处理一些较为复杂的模型。除此之外，实现映射能力、记忆机制与容错性也更为优良。因此本文考虑利用神经网络来优化情感分析模型。

1 文本的情感分析

■ 1.1 神经网络模型

神经网络是机器学习领域重要技术，能够较好的模拟人脑，实现人工智能。和人类脑神经网络结构类似，人工神经网络具有能够拟合出复杂模型的特点，图 1 是神经网络模型，其主要包含了三层结构，首先是输入层，负责输入信号；其次是隐藏层，针对输入信号进行处理；最后是输出层，输出相关内容。对于有更多层数的神经网络来说，中间的所有层都称之为隐藏层。它连接了输入层与输出层，并通过最后一个隐藏层将数值传输给输出层。

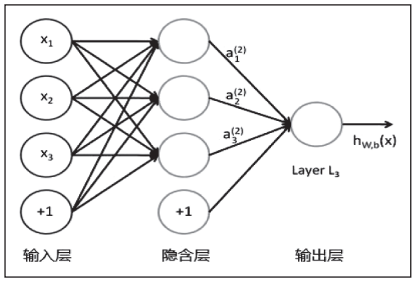

图 1　神经网络模型

设计好一个神经网络模型后，通常不能确认其是否高效准确，这时就需要这三步分别为：训练、开发、测试。首先，需要大量的数据来训练它，争取使这个模型达到最优，这些数据称为训练集，训练的过程是全自动的。神经网络模型可以设计成三层、五层等结构，但每个模型最终达到的准确度是不同的，比如用一个三层模型、五层模型、十层模型，用开发集中的数据对它们进行测试，开发就是找出准确度最高的模型。测试集是体现模型的泛化能力，例如老师交给学生函数知识，老师不可能把所有函数题型都教给你，所以学生会的函数问题越多，学生的泛化能力就越强。

■ 1.2 前向传播算法

前向传播算法[2]是神经网络最基本的算法之一，由第一层神经元传递向第二层时，假如 X 为输入，共有三个分量，计算的过程是如下公式 1 所示，θ 称权重，x_1、x_2　x_3 分别是上一层的三个输入，x_0 表示的是正 1，作为偏置项，同样的 a_1、a_2　a_3 表示下一层的三个神经元，a_0 表示的是正 1，作为偏置项，因为第二层的任意一个神经元是由第一层所有神经元通过乘权重加偏置得到的，如果所有的权重都相同，那么就会导致下一层所有的神经元都没有任何区别，导致对称失衡，所以我们给权重赋值时要随机赋予一个初始值，这个过程叫初始化。

$$a_1^{(2)}=\sigma\left(\theta_{10}^{(1)}x_0+\theta_{11}^{(1)}x_1+\theta_{12}^{(1)}x_2+\theta_{13}^{(1)}x_3\right)$$

$$a_2^{(2)}=\sigma\left(\theta_{20}^{(1)}x_0+\theta_{21}^{(1)}x_1+\theta_{22}^{(1)}x_2+\theta_{23}^{(1)}x_3\right)$$

$$a_3^{(2)}=\sigma\left(\theta_{30}^{(1)}x_0+\theta_{31}^{(1)}x_1+\theta_{32}^{(1)}x_2+\theta_{33}^{(1)}x_3\right)$$

$$h_{\theta(x)}=\sigma\left(\theta_{10}^{(2)}a_0^{(2)}+\theta_{11}^{(2)}a_1^{(2)}+\theta_{12}^{(2)}a_2^{(2)}+\theta_{13}^{(2)}a_3^{(2)}\right) \quad (1)$$

因为加权的运算属于线性运算，多次的线性运算和单次没有区别，这样就会导致多层的神经网络的复杂性无法体现，因此在每做完一次加权运算之后，都需用激活函数做非线性映射，如公式 2 为几个常见的激活函数，激活函数一方面是为了体现非线性的特点，另一方面它们求导也非常方便，例如 sigmoid 函数的导数就等于它自身乘上一减去它自身，这为反向传播的计算可以简化计算量，下面列举了几种激活函数的求导值，可以看出其计算简便。

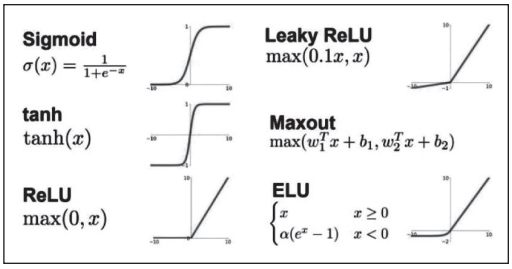

图 2　激活函数

$$\sigma'(z)=\sigma(z)\left(1-\sigma(z)\right)$$
$$tanh'(z)=1-\left(tanh(z)\right)^2 \quad (2)$$
$$\mathrm{Re}lu'(z)=\begin{cases}1\ when\ z\geqslant 0;\\ 0\ when\ z<0\end{cases}$$

■ 1.3 反向传播算法

在通过神经网络的计算得到一个从输入映射的输出值之后，为了评判模型的好坏，提出了损失函数用来衡量其与真实值之间的差距，损失函数也称为代价函数。通过使用损失函数来计算误差值，并反复计算更新权值，直至误差值最小。常用损失函数公式为：

$$J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}\mathrm{Cos}t\left(h_\theta\left(x^{(i)}\right),y^{(i)}\right)$$
$$\mathrm{Cos}t\left(h_\theta\left(x^{(i)}\right),y^{(i)}\right)=\begin{cases}-\log\left(h_\theta(x)\right) & if\ y=1\\ -\log\left(1-h_\theta(x)\right) & if\ y=0\end{cases} \quad (3)$$

其中， y=1 时， h(Θ) 为 1，误差为 0； h(Θ) 不为 1，误差与 h(Θ) 呈现负相关。y=0 时，h(Θ) 为 0，误差为 0；h(Θ) 不为 0，误差与 h(Θ) 呈现正相关。

反向传播算法是目前神经网络领域常用的有效算法，它利用的就是损失函数以及梯度下降算法的使用。梯度下降算法如公式 4 所示。

$$\theta_j:=\theta_j-\alpha\frac{\partial}{\partial\theta_j}J(\theta) \quad (4)$$

其中 α 为学习速率即步长，它的值越小，梯度下降就越慢；而越大可能导致无法收敛，所以 α 的选择很重要，只有合适的学习速率才能够找到极值点。在这里，反向传播算法不是一个具体的算法，而是对梯度下降算法的描述，通过求损失函数关于所有参数的偏导数作向前传播。首先计算损失函数关于最后一层参数的偏导数，然后利用该偏导数去用梯度下降算法更新参数，再一层层反向求出各层更新后的参数，直到所有参数都被更新。

■ 1.4 文本情感分析方法

设计一个模型，首先要人工标注，对于不同的话语，标注上积极与消极的两类情感分析，相当于在每一句话上贴上一个标签，其次再进行数据处理，把单词或文字转化为向量，转化为计算机认识的数字，接下来进一步对数据进行再处理，因为神经网络模型输入要求长度一致，所以需要把较长的数据进行分割，使得所有的数据信息基本长度一致；再其次是相关的模型设计，输入层通过前向传播算法到输出层，每个输入的神经元乘以相应的权重，再利用激活函数进行非线性的映射，得到下一层一个神经元的数值。通过相同的运算，只是使用不同的权重和偏置，可以得到下一层所有神经元的数值，这样就获得了神经元网络模型第一层到第二层的传递过程。然后使用相同的传递方式逐步计算到最后一层。建立好模型之后再进行模型评估，检查模型泛化能力的情况，最后做预测，用训练好的神经网络对新数据进行预测。

■ 1.5 实验及结果

（1）相关软件

此次实验我主要使用了 Python 语言，因为 Python 是一款跨平台的程序设计语言，常用于人工智能、软件开发、图像处理等领域。它具有简单、易学，速度快、免费开源等优点，对于初学者非常友好，通过学习容易上手。另外使用了 TensorFlow 库，它是谷歌开发和维护的一款系统，常应用于机器学习领域。

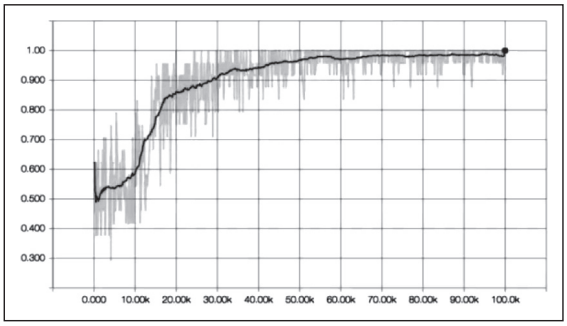

图 3　模型准确率曲线

（2）模型训练

考虑到数据来源，我利用了 NLPCC2013 的开放数据集进行训练，识别微博关键句中的评价对象和极性。训练数据由两个微博主题组成，每个主题各一百条，内含标注及数据

（下转第 43 页）

断向量号在向量表中的四个单元中，本例分配给 IR0 的中断类型号是 60H。第二，8259 初始化及相关设置，即对 8259 的工作方式等信息进行编程，依次给 8259 的偶奇端口送初始化命令字及操作命令字。第三，等待中断，本例的中断是由按键按下产生的，什么时候按下按键不确定，在主程序中可以安排一条原地跳指令即可实现等待中断过程，如: L:JMP L。第四，中断子程序，本例的中断子程序主要完成的工作是当有中断产生时输出数据到锁存器，控制发光二级管亮，由于每次只点亮一个共阴极的发光二极管，输出的数据里面有一位是 1 即可，可以采用循环指令来实现。

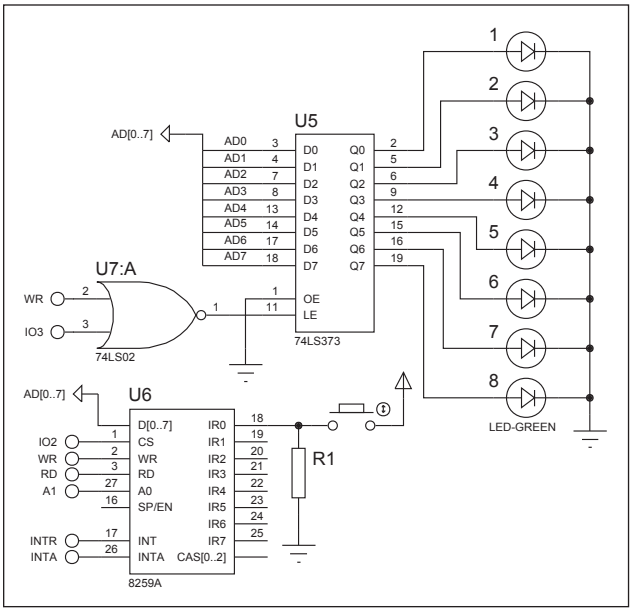

图 2　8259 中断控制器应用实例

对实践性很强的课程来说，“实践过程验证”法是一个非常有效的方法，它可以直接通过具体的应用实例，对所学的内容进行全面的加强和巩固，是电子信息工程类学生提升自身能力的必要途径。

2 结论

本文以目前用人单位对毕业生综合能力要求严苛为背景，提出高校教师应及时进行课程教学改革，以培养知识扎实，能综合处理问题的高素质人才为目标，以适应社会快速发展的需要。作者结合自己多年《微机原理与接口技术》课程的教学经验，积极探研，总结出三种实用且高效的教学方法，即：“新旧知识关联”法、“抽象概念实体化”法和“实践过程验证”法。经学院多个教学班中推广使用，效果令人满意，不仅提高了学生的学习兴趣，还教授了学生新的学习方法，使其知识技能能力全面提高，能更好地应对日益激烈的竞争环境。

参考文献

* [1] 郭晶晶 , 刘伯运 , 梁英杰 , 史蓓蕾 . 基于虚拟仪器技术的《微机原理与接口技术》教学改革探索 [J]. 教育教学论坛 , 2019(15) : 134-136.
* [2] 汤书森 , 段东波 , 高国平 , 张北斗“新工科”背景下电子信息类专业的实验教学改革实践——以微机原理与接口技术实验课为例 [J]. 高校实验室科学技术 , 2019(01) : 21-23.
* [3] 肖洁 , 洪连环 , 方平 . 基于 Proteus 仿真的《微机原理及应用》实验教学改革与实践 [J]. 软件 , 2019, 40(02) : 59-62.
* [4] 薛娓娓 , 李娣娜 , 王巧霞 . 基于 Proteus 的“微机原理与接口技术”实验教学研究 [J]. 无线互联科技 , 2017(23) : 90-92.
* [5] 张华 , 邹小花 , 王海威 , 李华英 , 马丁 . 独立学院电类专业微机原理与单片机课程群的教学改革与实践 [J]. 中国新通信 , 2016, 18(23) : 129.

（上接第 35 页）

说明。利用好 word2vec 将样本转化为词向量之后，就可进行模型训练。

（3）实验结果

通过实验得出，当模型训练次数达到 6 万次左右时，此时的准确率较高，约为 96%，这说明我的模型取得了很好的效果。

2 总结

本文通过对神经网络的研究，了解了神经网络的基本原理和方法，和其在文本情感分析上的应用，利用 Python 语言和 TensorFlow 库进行了实验验证，并取得了理想效果，验证了模型的有效性。未来工作拟从以下几方面开展，例如研究更加复杂的神经网络模型，以及设计神经网络结构，这都是接下来需要解决的问题。

参考文献

* [1] 殷昊 . 面向微博文本的情绪识别和分类方法研究 [D]. 苏州大学 ,2018.
* [2] 李然 , 林政 , 林海伦 , 王伟平 , 孟丹 . 文本情绪分析综述 [J]. 计算机研究与发展 ,2018,55(01):30-52.
* [3] 胡健楠 . 中文文本情绪原因发现研究 [D]. 哈尔滨工业大学 ,2018.