基于递归神经网络和快速文本的社交大数据分析

孙二华

（重庆建筑科技职业学院信息工程学院，重庆 401331）

摘　要：针对传统模型无法有效地从大数据中提取和分析相关信息等问题，提出了基于递归神经网络和快速文本的社交大数据分析方法。该方法首先对输入数据进行预处理以减少噪声，然后基于快速文本词嵌入将文本数据转换成数值向量，基于递归神经网络学习每个输入序列的最佳特征表示，最后基于 Logistic 回归算法使用最佳特征表示执行分类任务。实验结果表明，与其他方法相比，在分类准确性和 F 得分方面均优于对比的方法，能够处理大规模数据和提高递归神经网络变体在分类精度方面的性能。

关键词：递归神经网络；大数据；长短期记忆网络；门控循环单元

DOI:10.16184/j.cnki.comprg.2020.09.040

1　概述

社交媒体产生的大数据代表着宝贵的信息来源，通过社交媒体上的评论了解用户对购买的产品或使用服务的看法是做出决策的关键因素[1, 2]，如何高效地挖掘大数据成为当前研究的热点。大数据分析是指通过各种解决方案和技术分析海量数据以获得有意义的见解，用于预测、分类和决策等目的[3, 4]。除了分析海量数据外，大数据分析还为机器学习技术和数据分析任务带来了噪声数据、高度分散的输入数据源、高维、有限的标签数据等严峻的挑战，还存在数据索引、数据存储和信息检索其他实际问题。因此，需要更复杂的数据分析和数据管理工具来处理大数据环境下的海量数据和各种现实问题[5, 6]。

深度学习模型在包括文本分类和情感分析在内的各种自然语言处理任务中取得了显著成功，用于自动学习分层表示并基于深层架构提取高级特征[7]。长短期记忆（Long Short-Term Memory, LSTM）是一种具有反馈连接的递归神经网络（Recurrent Neural Network, RNN），包括一个存储单元和 3 个调节器或门（输入门、输出门和忘记门），用于控制 LSTM 单元内部的信息流[8, 9]。存储单元保持输入特征之间的依赖性，输入门将新值输入到存储单元中，忘记门控制值是否保留在存储单元中，输出门使用存储单元中的值计算单元的输出，切线函数和 S 型函数是 LSTM 单元的常见激活函数[10]。

2　递归神经网络模型（RNN）

随着深度学习的成功，基于递归神经网络（RNN）模型将单词转换为有意义的向量越来越受欢迎。LSTM 是一种特殊类型的递归神经网络，能够克服消失的梯度，并且比标准的递归神经网络有更好的性能。设 N 为 LSTM 单元数量，一个 LSTM 单元具有隐藏状态 h_t、存储单元 c_t 和 3 个门（输入门 i_t、忘记门 f_t 和输出门 o_t）。每个门采用 S 型函数作为激活函数，并产生区间 [0,1] 中的值。

$$\begin{aligned} i_t &= \sigma\left(W_i v_t + U_i h_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_f v_t + U_f h_{t-1} + b_f\right) \\ o_t &= \sigma\left(W_o v_t + U_o h_{t-1} + b_o\right) \end{aligned} \tag{1}$$

其中 σ 表示 S 形函数。W 和 U 是门的权重矩阵，b 是偏移权重。在时间步长 t，词表示向量 $v_t \in \mathbb{R}^D$、先前的隐藏状态 $h_{t-1} \in \mathbb{R}^M$ 和先前的存储单元向量 c_{t-1} 作为 LSTM 的输入。

$$\begin{aligned} g_t &= \tan h\left(W_g v_t + U_g h_{t-1} + b_g\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tan h\left(c_t\right) \end{aligned} \tag{2}$$

其中符号⊙为逐个素相乘，W_g 和 U_g 是权矩阵，b_g 为偏差项。BiLSTM 拥有两个 LSTM，第一个是由 $\overrightarrow{LSTM}$ 表示从左到右处理输入序列的前向 LSTM，第二个是由 $\overleftarrow{LSTM}$ 表示从右到左处理输入序列的反向 LSTM。在每个

基金项目：重庆市高等教育教学改革研究项目，（编号：202156）；重庆市教委高职教育双基地建设项目（编号：20180310）。

作者简介：孙二华（1978–），女，硕士，副教授，研究方向：大数据、物联网。

时间步长 t，通过连接前向隐藏状态$\vec{h}_t$和后向隐藏状态$\overleftarrow{h}_t$来生成隐藏状态[13]。

$$\vec{h}_t = L\vec{S}TM(v_t, \vec{h}_{t-1})$$
$$\overleftarrow{h}_t = L\overleftarrow{S}TM(v_t, \overleftarrow{h}_{t+1}) \quad (3)$$

其中$\vec{h}_{t-1}$为先前隐藏状态，$\overleftarrow{h}_{t+1}$为未来隐藏状态，v_t为输入序列。门控循环单元网络（Gated Recurrent Unit, GRU）通过减少每个单元中可训练参数的数量来简化LSTM网络的复杂体系结构。它只有复位门 r_t 和更新门 u_t 两个用于控制隐藏状态的门，在时间步长 t，GRU 的激活 h_t 和候选激活 $\hat{h}_t$ 计算如下所示：

$$r_t = \sigma\left(W_r v_t + U_r h_{t-1} + b_r\right)$$
$$u_t = \sigma\left(W_u v_t + U_u h_{t-1} + b_u\right) \quad (4)$$
$$\hat{h}_t = \tan h\left(W_{\hat{h}} v_t + U_{\hat{h}}\left(r_t \odot h_{t-1}\right) + b_{\hat{h}}\right)$$
$$h_t = u_t \hat{h}_t + \left(1 - u_t\right) \odot h_{t-1} \quad (5)$$

其中 W 和 U 是权重矩阵，b 是偏差项。

3 文本预处理

文本预处理是文本分析的重要环节，是为文本分析准备数据的过程。评论通常由嘈杂和结构不良的句子组成，因此有必要将每次评论的内容转换为适当的格式，以减少噪音的负面影响并促进文本分析。执行数据清理步骤通过应用平滑技术来消除或减少噪声，消除异常值。使用标准归一化技术执行数据转换和还原步骤，有助于提供更容易理解的预测模式，执行相关性分析以检测对预测任务没有贡献的冗余属性。

快速文本是一种最流行的基于神经网络的单词嵌入技术，是专门为在考虑语法的同时学习单词的高质量表示而设计的。一般来说，FastText 是基于连续 Skipgram 模型和子词模型的。

负抽样 Skipgram 模型：设 W 是单词词汇量，其中每个单词由其索引$w \in \{1, \cdots, W\}$标识，给定训练语料库定义为单词 $w_1, \cdots w_r$ 的序列，Skipgram 算法的主要目标是最大化以下目标函数：

$$O = \sum_{t=1}^{T} \sum_{c \in C_t} \log\left(p\left(w_c \mid w_t\right)\right) \quad (6)$$

其中 C_t 为词 w_t 周围的词索引集合，上下文词的预测问题可以视为一组独立的二进制分类任务，目的是预测上下文词的存在或不存在。对于位置 t 处的词，所有上下文单词均被视为正例，而反例是从词汇表中随机抽取的。因此目标函数定义如下：

$$\log\left(1 + e^{-s(w_t, w_c)}\right) + \sum_{n \in N_{t,c}} \log\left(1 + e^{s(w_t, n)}\right) \quad (7)$$

$$s\left(w_t, w_c\right) = u_{wt}^T v_{wc} \quad (8)$$

其中 $N_{t,c}$ 为从词汇表中抽取的一组反例。设 u_{wt} 和 v_{wc} 分别对应于单词 w_t 和 w_c 的两个向量，函数$s(w_t, w_c)$为单词和上下文向量之间的标量积。

子词模型：将每个词被视为字符 n-gram 组成的包。设 D 为大小为|D|的 n-gram 字典，对每个词 w，D_w 表示出现在 w 中的 n-gram 集合，其中每个矢量 z_d 与表示为 d 的每个 n-gram 相关联。因此，每个词 w 由其 n-gram 的向量表示之和表示。该模型有助于跨单词共享表示形式，从而学习稀有单词的可靠表示形式。

$$s(w, c) = \sum_{d \in D_w} z_d^T v_c \quad (9)$$

设$R^{(i)} = (w_1, \cdots w_N)$为输入序列，表示填充后 N 个词组成的评论。为了处理基于递归神经网络变体的评论$R^{(i)}$，将文本数据转换成数值向量，每个向量捕捉有关语言的隐藏信息。基于预训练的快速文本词表示形式，数据集中的每个评论$R^{(i)}$可以转换为矢量序列：

$$V^{(i)} = FastText\left(R^{(i)}\right) = (v_1, \cdots, v_N) \quad (10)$$

其中每个单词 w_i 由 D 维向量$v_i \in \mathbb{R}^D$表示。

4 文本分类

文字分类是指从具有预定义标签的评论中学习分类模型，然后将学习到的模型用于将来的评论分类。本文采用 Logistic 回归算法进行分类任务，该算法是一种有监督的学习方法。在表示过程之后，训练集可以表示如下：

$$T = \left\{\left(x_i, y_i\right), i = 1, \cdots, |T|\right\}, x_i \in \mathbb{R}^k \quad (11)$$

其中$x_i = O^{(i)}$为推算的表示形式（即特征），y_i 是与评论$R^{(i)}$相关的类别标签，目的是训练可以将向量$O^{(i)} \in \mathbb{R}^k$作为输入的多项式 Logistic 回归模型，并预测不同可能类别的概率。使用训练的模型，x_i 属于类别标签 y_i 的概率：

$$P\left(y_i^{(c)} = 1 \mid x_i; \theta\right) = \frac{\exp\left(\theta^{(c)^T} x_i\right)}{\sum_{j=1}^{C} \exp\left(\theta^{(j)^T} x_i\right)} \quad (12)$$

其中$c \in \{1, \cdots, C\}$，如果类别标签是 c，则$y_i^{(c)}$为 1，否则为 0。$\theta^{(c)}$表示与类 c 相对应的参数向量，T 表示矩阵转置，θ表示模型的所有参数。最后基于估计的概率，模

型生成预测类。

5 实验结果与分析

所有实验均在分布式大数据分析平台上进行，该平台由一个主节点和一个从节点实现，单个节点在具有Intel(R)Core(TM)i7-6500U 处理器和 8.00 GB 内存的机器上执行，采用 Python3.5 编写深度学习模型，采用 FastText 进行分布式单词表示。选择两个真实世界的数据集 Yelp 和 Sentiment140 进行实验。Yelp 数据集由 1637138 名用户为 192609 家企业提供的 6685，900 条分类评论组成。本文随机选取了 10 万条评论作为原始数据集，认定 1 星和 2 星是负面的，而 4 星和 5 星是正面的。Sentiment140 是一个来源于斯坦福大学的推特情绪分析数据集，由 160 万条分类推文组成，随机选取 20000 条推文作为原始数据集。

使用 Adam 优化算法对不同的深度学习模型进行优化，进行大量实验确定最佳设置。批处理大小设置为 60，密集层的 dropout 率固定为 0.5，而复发性 dropout 率固定为 0.2，表示参数 k 设置为 9，学习率固定为 0.01。

6 结语

为了处理社交大数据并提高其分类准确性，提出了基于递归神经网络变体和快速文本词嵌入的社交大数据分析方法。该方法的主要目的为处理大规模数据和提高分布式深度学习模型的分类性能，它首先对数据进行预处理以将评论的内容转换成适当的格式，然后基于快速文本词嵌入将文书数据转换为数值向量，其中每个向量捕获有关语言的隐藏信息，基于分布式 RNN 变体学习每个输入序列的最佳特征表示。实验结果表明，提出的方案可以提高深度学习模型长短期记忆（LSTM），双向长短期记忆（BiLSTM）和门控递归单元（GRU）的分类性能，有助于更好地理解公众舆论和用户行为。

参考文献

[1] 孙倩，陈昊，李超. 基于改进人工蜂群算法与 MapReduce 的大数据聚类算法 [J] . 计算机应用研究，2020，37 (06).

[2] Oussous A, Benjelloun F Z, Lahcen A A, et al. Big Data technologies: A survey [J] . Journal of King Saud University-Computer and Information Sciences, 2018, 30 (4) : 431-448.

[3] Sivarajah U, Kamal M M, Irani Z, et al. Critical analysis of Big Data challenges and analytical methods [J] . Journal of Business Research, 2017, 70: 263-286.

[4] Gunther W A, Mehrizi M H R, Huysman M, et al. Debating big data: A literature review on realizing value from big data [J] . The Journal of Strategic Information Systems, 2017, 26 (3) : 191-209.

[5] Jimenez-Marquez J L, Gonzalez-Carrasco I, Lopez-Cuadrado J L, et al. Towards a big data framework for analyzing social media content [J] . International Journal of Information Management, 2019, 44: 1-12.

(上接第 75 页)

技术进行连接，实现了硬件虚拟化，可以扩展地理位置服务保存的服务器数量，从而可以提高存储空间利用率。同时，地理位置服务保存的数据很多都是卫星地图，这些地图的数据量非常大，因此常规的服务器需要购置大规模的设备，花费的资金比较多，同时用户实时访问卫星地图也需要很大的带宽和并发计算能力。云计算能够为用户提供一个强大的并发访问接口，这个接口可以帮助用户实现信息服务，提高信息加工和保存能力。

4 结语

移动地理位置服务可以利用智能手机、电子地图和导航卫星等进行融合定位服务，为用户提供强大的交通导航、旅游景点线路规划、酒店位置查询等功能，还可以实现物流运输目标的定位，提高人们工作、生活和学习的信息化。

参考文献

[1] 王磊，孙中伟. 基于安全网络编码的移动网络地理位置隐私保护技术 [J] . 南京理工大学学报，2018，218 (01) : 60-65.

[2] 王兴国. 基于 Android 平台的移动 GIS 旅游信息服务应用研究 [J] . 度假旅游，2018，24 (02) : 129-131.

[3] 乐洪舟，张玉清. 网络直播平台主播地理位置泄露漏洞的分析与利用 [J] . 计算机学报，2019，042 (005) : 1095-1111.

[4] 姚登敏，李百寿，沈宇臻. 面向智慧景区位置服务的通信定位性能空间优化分析 [J] . 中国科技信息，2018，(17) : 97-100.