基于深度学习的文本特征提取研究综述

张　千,王庆玮,张　悦,纪校锋,张宇翔,祝　赫,赵昌志
(中国石油大学(华东) 计算机与通信工程学院,山东 青岛 266580)

摘　要:文本特征项的选择是文本挖掘和信息检索的基础和重要内容。传统的特征提取方法需要手工制作的特征,而手工设计有效的特征是一个漫长的过程,但针对新的应用深度学习能够快速地从训练数据中获取新的有效特征表示。作为一种新的特征提取方法,深度学习在文本挖掘方面取得了一定的成果。深度学习与传统方法的主要区别在于,深度学习能自动地从大数据中学习特征而不是采用手工制作的特征,手工制作的特征主要依赖于设计者的先验知识,很难充分利用大数据;深度学习可以自动地从大数据中学习特征表示,并包括数以万计的参数。文中概述了用于文本特征提取的常用方法,并阐述了在文本特征提取及应用中常用的深度学习方法,以及深度学习在特征提取中的应用展望。

关键词:深度学习;特征提取;文本特征;自然语言处理;文本挖掘
中图分类号:TP31　　**文献标识码**:A　　**文章编号**:1673-629X(2019)12-0061-05
doi:10.3969/j.issn.1673-629X.2019.12.011

Review of Text Feature Extraction Based on Deep Learning

ZHANG Qian, WANG Qing-wei, ZHANG Yue, JI Xiao-feng, ZHANG Yu-xiang,
ZHU He, ZHAO Chang-zhi
(School of Computer & Communication Engineering, China University of Petroleum (East China), Qingdao 266580, China)

Abstract: The selection of text feature items is basic and important in text mining and information retrieval. Traditional feature extraction methods require hand-made features, and manual design of effective features is a long process. However, for new applications, deep learning can quickly obtain new and effective feature representation from training data. As a new feature extraction method, deep learning has made some achievements in text mining. The main difference between deep learning and traditional methods is deep learning can automatically learn features from large data rather than using hand-made features. Hand-made features mainly rely on designer's prior knowledge, which is difficult to fully use large data. Deep learning can automatically learn feature representation from large data and include tens of thousands of parameters. We summarize the common methods of text feature extraction and expound the deep learning methods commonly used in text feature extraction and application, as well as the application prospect of depth learning in feature extraction.

Key words: deep learning; feature extraction; text characteristic; natural language processing; text mining

0　引　言

机器学习是人工智能的一个分支,在许多情况下几乎成了人工智能的代名词。机器学习系统用于识别图像中的对象,将语音转换成文本,匹配用户感兴趣的新闻、文章或产品,并选择相关的搜索结果[1]。这些应用程序越来越多地使用了一种叫做深度学习的技术,而传统的机器学习技术在以原始的形式处理自然数据的能力上受到了限制[1-2]。

几十年来,构建一种模式识别或机器学习系统需要周密的工程和相当大的专业领域知识。设计一种特征提取方法,将原始数据(如图像的像素值)转化到一个合适的内部特征向量或表现形式。学习子系统往往是一个分类器,可以检测或辨别输入模式分类[1];表示学习是一组方法,它允许机器对原始数据进行反馈,并自动发现用于检测或分类需求的表示[1]。深度学习方法是通过组合简单而非线性的模块而获得的有着多层次表现的表示学习方法,每个模块从一个层次(从原始输入)转换到一个更高、更抽象的层次表示,由于有足够的这种变换故可以学习到相对复杂的函数[1-2]。

文本特征提取是一个从文本信息提取到展现文本

收稿日期:2018-10-31　　**修回日期**:2019-03-06　　**网络出版时间**:2019-09-24
基金项目:中央高校基本科研业务专项基金(18CX02019A);科技部创新方法工作专项(2015IM010300)
作者简介:张　千(1982-),女,副教授,研究方向为大数据智能处理、智慧医疗;王庆玮(1996-),女,在读硕士,研究方向为大数据智能处理、智慧医疗。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1534.006.html

信息的过程,是进行大量文本处理的基础[3-4]。在特征提取过程中,将删除不相关或多余的特征。特征提取作为学习算法的一种数据预处理方法,能更好地提高学习算法的精度并节省时间。常用的文本特征提取方法有过滤、融合、映射和聚类等。深度学习的关键在于这些特征层是不需要人设计的,而是使用通用学习程序从数据中学习[1]。深度学习只需要很少的手工量,因此可以很容易地利用现有计算和数据量的增加[1]。深度学习善于识别非结构化数据的模型和大多数熟悉的媒体,如图像、声音、视频、文本等。目前,深度学习的特征表示包括自编码、限制 Boltzmann 模型、深度信念网络、卷积神经网络和递归神经网络等。

1 文本特征提取方法

文本特征提取在文本分类中起着重要的作用,能直接影响文本分类的准确率[3-5]。它是基于向量空间模型(VSM),其中文本被看作是 n 维空间中的一个点,点的每个维度的数据代表文本的一个数字化特征。文本特征通常使用关键字集,它是指在一组预定义关键词的基础上,用一定的方法计算文本中词的权重,然后形成一个数字向量,即文本的特征向量。现有的文本特征提取方法包括过滤、融合、映射和聚类方法等。

1.1 过滤方法

过滤速度快,特别适用于大规模文本特征提取,过滤文本特征提取主要有词频、信息增益和互信息法等。

1.1.1 词 频

词频是指一个词出现在文本中的次数。通过词频特征选择,即删除频率小于某一阈值的词,以减少特征空间的维数。这种方法基于这样一个假设:小频率的单词对过滤的影响很小[3,6-7],而在信息检索的研究中,人们认为有时频率较低的词可能会包含更多的信息。因此,在特征选择过程中,仅仅基于词频来删除大量的词汇是不合适的。

1.1.2 互信息

用于计算两个对象相互度量的互信息法(互信息,MI)[8-9]是计算语言学模型分析中常用的方法,用于测量在过滤中从特征到主题的区别。互信息的定义类似于交叉熵。对于互信息理论进行特征提取是基于如下假设:在某一类中有较大词频的单词但在其他类中词频较小,且类具有较大的互信息。通常互信息被用作特征词和类之间的度量,如果特征词属于类,则它们拥有最大数量的互信息。由于该方法不需要对特征词与类之间的关系进行任何假设,因此非常适合于文本分类和类特征的注册[9]。

1.1.3 信息增益

IG(信息增益)是机器学习的常用方法。在过滤中,它被用来衡量一个已知特征是否出现在某个相关主题的文本中,以及该主题的预测信息有多少。信息增益是一种基于熵的评价方法,涉及到大量的数学理论和复杂的熵理论公式。它定义为某个特征项能够为整个分类提供的信息量,不考虑特征的熵而是特征熵的差值[10]。根据训练数据计算每个特征项的信息增益,并删除基于信息增益的小信息项,其余部分按信息增益降序排列。

1.1.4 应 用

文献[11]中提出一种基于特征聚类算法的无监督特征提取方法,它对利用互信息最大化(MIM)方法寻找合适的聚类特征变换进行了研究。UCI 数据集的实验表明,该方法在分类精度方面优于传统的无监督方法 PCA(主成分分析)和 CA(独立分量分析);文献[12]中,针对传统词频索引逆文档频率提取算法(TF-IDF)效率低、准确性差的问题,提出了一种基于词频统计的文本关键词提取方法。实验结果表明,在关键词提取的查准率、查全率等指标方面,基于词频统计的 TF-IDF 算法均优于传统的 TF-IDF 算法,且能有效降低关键词提取的运行时间;在参考文献[13]中,提出一种特征选择的组合方法,该方法将基于相关的滤波器应用于整个特征集以寻找相关的特征,然后在这些特征上应用包装器,以找到指定预测器的最佳特征子集。

1.2 融合方法

融合需要特定分类器的集成,在指数增长区间内进行搜索,这种方法时间复杂度高,因此不适用于大规模的文本特征提取。加权法是一种特殊的融合方法,在[0,1]以内的每个特征权重都将进行训练并进行调整。线性分类器集成的加权方法是高效的,KNN 算法是一种基于实例的学习方法[14]。

1.2.1 加权 KNN(k 最近邻)

Han[15]提出了一种结合 KNN 分类器的加权特征提取方法,该方法能将每个连续累积值进行分类并具有良好的分类效果。KNN 方法作为一种基于统计模式识别的无参数文本分类方法,能得到较高的分类准确率和查全率[14-15]。

1.2.2 中心向量加权法

Shankar 提出加权中心向量分类法,先定义一种具有区分能力的特征方法,然后利用这种能力有权区分新的中心向量,算法需要多重加权直到分类能力下降。

1.3 映射方法

映射广泛应用于文本分类并取得了良好的效果,它通常用于 LSI(潜在语义索引)和 PCA 中。

1.3.1 潜在语义分析

LSA(或 LSI)是一种新型信息检索代数模型,是

用于知识获取和演示的计算理论或方法,采用统计计算的方法对大量文本集进行分析,提取词间潜在的语义结构,利用这种潜在的结构来表示词和文本,从而通过简化文本向量消除词之间的相关性并减少维数[10]。

潜在语义分析的基本概念是将高维 VSM 中的文本映射到低维潜在语义空间,这种映射是通过项目或文档矩阵的 SVD(奇异值分解)来实现的[14]。

LSA 的应用:信息过滤、文档索引、视频检索、文本分类与聚类、图像检索、信息提取等。

1.3.2　最小二乘映射方法

Jeno 对基于中心向量和最小二乘法的高维数据约简做了研究,他认为由于聚类中心向量反映了原始数据的结构而 SVD 不考虑这些结构,所以降维比 SVD 更具有优势。

1.3.3　应　用

文献[16]中提出了一种新的滤波器,这种滤波器基于盖然论的概率特征选择方法,即 DFS(基于特征选择)文本分类方法。实验对不同的数据集、分类算法和成功措施进行了比较,结果表明 DFS 在分类精度、降维率和处理时间方面提供了有竞争力的性能[16]。

1.4　聚类方法

聚类法考虑到文本特征的本质相似性,主要是对文本特征进行聚类,然后使用每个类的中心来替换该类的特性。该方法压缩比很低并且分类精度基本保持不变,但是复杂度较高。

1.4.1　CHI(卡方)聚类法

通过每个特征词对每个类(每个特征词得到对每个类的 CHI 值)贡献的计算,CHI 聚类法聚类文本特征词对分类的相同贡献,使它们的共同分类模型取代了传统算法中每个单词对应一维的模式。

1.4.2　概念索引

在文本分类中,概念索引(CI)是一种简单有效的降维方法。通过将每个类的中心作为基向量结构的子空间(CI 子空间),然后将每个文本向量映射到这个子空间,得到文本向量到子空间的表示。训练集所包含的分类量正是 CI 子空间的维数,通常小于文本向量空间的维数,从而实现向量空间的降维。

1.4.3　应　用

文献[17]对利用遗传算法和模糊聚类技术将大特征空间与有效数字相结合的两种方法进行了描述,最后利用自适应神经模糊技术实现了模式的分类。整个工作的目的是实现对人脑肿瘤病变分类的识别,即 CT 和 MR 图像所确定的占位性病变。

2　深度学习方式

深度学习是在 2006 年由 Hinton 等提出的一类无监督学习,它的概念来源于人工神经网络的研究。深度学习结合底层特征形成更抽象、更高层次的属性表征或特征,深层次地发现数据的分布特征表示[2]。

深度学习与表面学习相反,现在很多学习方法都是表面结构算法,而且它们都存在一定的局限,如在有限样本的情况下复杂功能性具有局限,对复杂分类问题的泛化能力受到一定的限制[18]。深度学习和传统的模式识别方法间的主要区别是深度学习能够自动地从大数据中学习特征,而不是采用手工制作的特征[2]。在计算机视觉发展史上,五年到十年才能出现一个被广泛认可的优良特性,但是针对新的应用,深度学习能够快速从训练数据中获取新的有效特征表示。

深度学习技术应用在普通的 NLP(自然语言处理)任务中,如语义分析、信息检索、语义角色标注、情感分析、问答、机器翻译、文本分类、文本生成,以及信息提取。卷积神经网络和递归神经网络是常用的两种模型。接下来介绍文本特征提取的几种深度学习方法及其应用、改进方法和步骤。

2.1　自编码

自编码是一种前馈网络,可以学习数据的压缩分布式表示,通常以降维或流形学习为目标。自编码的隐藏层通常具有比输入层和输出层更紧凑的表示,它的单元比输入层或输出层要少。输入和输出层通常具有相同的设置,允许自编码进行无监督训练,即在输入端输入相同的数据,然后与输出层的数据进行比较。训练过程与传统的反向传播神经网络相同,唯一的区别在于通过输出与数据本身的比较来计算误差[2]。

堆叠的自编码是编码的深度对应,可以简单的通过堆积层建立。对于每一层,它的输入是前层的学习表示,可以学习到比现有学习更为紧凑的表示。文献[2]中针对短文本的特点,提出了特征提取和基于深度噪声的自编码聚类算法。该算法利用深度学习网络将高维、稀疏短文本的空间矢量转换为新的、低维的、实质性的特征空间。实验表明,将提取的文本特征应用于短文本聚类,显著提高了聚类效果。文献[2]中提出使用深度学习的稀疏编码自动提取文本特征,并结合深度信念网络形成 SD(标准差)算法的文本分类。实验表明,在训练集较少的情况下,SD 算法的分类性能比传统的支持向量机低,但在处理高维数据时,SD 算法比 SVM 算法具有更高的准确率和召回率;

2.2　受限玻尔兹曼机

RBM(受限玻尔兹曼机)于 1986 年由 Smolensky 提出,是玻尔兹曼机的可见单元之间或隐藏单元之间没有连接的受限版本[2]。该网络由可见单元(可见向量即数据样本)和一些隐藏单元(相应隐藏的向量)组成。有形载体和隐向量为二进制向量,即它们取{0,

1}之间的数值。整个系统是一个双向图,边缘只存在于可见单位和隐藏单元之间,可见单元之间和隐藏单元之间没有边缘连接(如图 1 所示)。

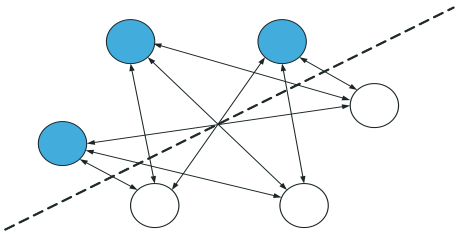

图 1 RBM

图 1 中,存在隐藏单元之间(阴影节点)没有连接而可见单元也没有连接(无阴影节点)的限制,Boltzmann 机变成一个 RBM。现在的模型是一个双向图。培训过程自动要求重复以下三个步骤:

(1)在正向传递过程中,每个输入与单个权重和偏置相结合,并将结果发送到隐藏层;

(2)在逆向过程中,每个激活与单个重量和偏置相结合,结果被传送到可见层进行重建;

(3)在可见层,利用 KL 散度对重建和初始输入进行比较,以决定结果质量。

使用不同的权重和偏差重复上述步骤,直到重建和输入尽可能接近为止。

2.3 深度信念网络

DBN(深度信念网络)是由 Hinton 等于 2006 年提出,他表明 RBMS 可以以贪婪的方式堆放和训练[2]。DBN 在网络结构方面都可以看作是一个堆栈,隐藏层中可见的受限玻尔兹曼机是该层上的一层。

经典 DBN 的网络结构是由一些 RBM 层和一层 BP 构成的深度神经网络。图 2 是三层 RBM 网络构成的 DBN 网络结构。DBN 的训练过程包括两步:第一步是分层预训练,第二步是 ne 调谐。

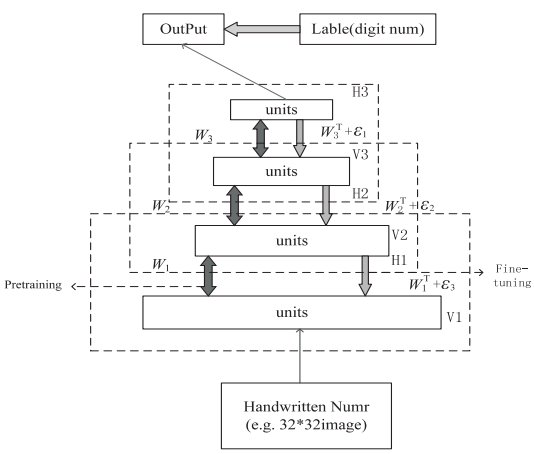

图 2 DBN 网络结构

DBN 模型的训练过程主要分为两个步骤:

(1)分别单独在没有监督下训练 RBM 网络各层,并且确保作为特征向量被映射到不同的特征空间,特征信息尽可能保留。

(2)在 DBN 的最后一层建立 BP 网络,将受限玻尔兹曼机的输出特征向量作为输入特征向量,并且在监督下训练实体关系分类器。每一层的 RBM 网络仅能确保自己层的量到该层的特征向量,而不是对整个 DBN 的特征向量进行优化。因此,反向传播网络传播自上而下的信息到每一层的 RBM,并微调整个 DBN 网络。RBM 网络训练模型的过程可以看作是一个深度的 BP 神经网络权值初始化的过程,能使 DBN 克服深度 BP 网络权重参数初始化导致的局部最优和长训练时间的缺点。

步骤(1)称为在深度学习术语中的预训练,步骤(2)称为微调。任何基于特定应用域的分类器在 BP 网络下都可以应用于有监督学习的层。

2.4 卷积神经网络

卷积神经网络(convolution neural network,CNN)是近年来发展起来的一种高效识别方法。CNN 网是一个多层神经网络,每一层都是由多个二维的表面组成,每个面是由多个独立的神经元组成。CNN 是人工神经网络的一种,具有较强的适应性,善于挖掘数据的局部特征。网络结构的权重使其更类似于生物神经网络,降低了网络模型的复杂度,减少了权值的数量,使 CNN 在模式识别的各个领域得到了应用,取得了很好的效果。CNN 结合本地感知区域,在时间或空间上降低采样来充分利用数据本身包含的诸如区特征之类的特征,并优化网络结构,保证一定程度的位移不变性。通过多年的研究,对神经网络的应用越来越多,如人脸检测、文件分析、语音检测、车牌识别等。2006 年,Kussul 提出将神经网络的置换编码技术应用于人脸识别、手写体数字识别和小目标识别技术中,这些技术通过分类系统的一些特殊性能来完成;2012 年,研究人员将视频数据中的连续帧作为神经网络输入数据的卷积,以便在时间维度上引入数据,从而识别人体运动。

2.5 递归神经网络

RNN 用来处理时序数据,在传统神经网络模型中,它从输入层到隐藏层再到输出层,这些层是完全连接的,并且每个层的节点之间没有连接。对于涉及顺序输入的任务,比如演讲和语言往往会更好地使用它[2](见图 3)。RNNs 一次一个元素地处理输入序列,在隐藏的单元中保持一个状态向量,隐含地包含关于序列所有过去元素的历史信息。当考虑隐藏单元在不同离散时间步长上的输出时,就好像它们是深度网络中不同神经元的输出,从而知道如何运用反向传播算法来训练网络[2]。

人工神经元(例如,时间 t 中的值 s_t 在节点 s 下分组的隐藏单元)在以前的时间步长中从其他神经元获

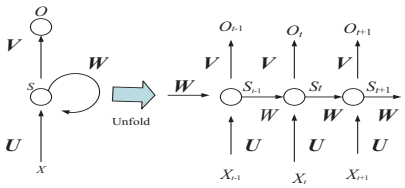

图 3　递归神经网络及其正向计算中计算时间的展开

得输入(这是用黑色方块表示的,表示在一个时间步长上的延迟)。这样,一个递归神经网络就可以将输入序列与 X_t 元素映射成一个带 O_t 元素的输出序列,其中每个元素 O_t 依赖于所有以前的 $X_{t'}$ (对于 $t' < t$)[2]。每个时间步长使用相同的参数(矩阵 $\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}$)。反向传播算法(图 1)可以直接应用于展开网络的计算图,计算所有状态 S_t 和所有参数的总误差(例如生成正确的输出序列的日志概率)的导数[2]。

3　结束语

文本特征项的选择是文本挖掘和信息检索的基础和重要内容。特征提取是指根据一定的特征提取指标,从测试集的初始特征集提取相关的原始特征子集,删除不相关或多余的特征,从而降低特征向量空间维度。特征提取作为学习算法的一种数据预处理方法,能更好地提高学习算法的精度,缩短学习时间。与其他机器学习方法相比,深度学习能从特征中检测复杂的相互作用,从几乎未处理的原始数据中学习低级特征,挖掘不易被检测到的特征,处理高基数的类成员和处理未开发的数据。与几个深度学习的模型相比,递归神经网络已广泛应用于自然语言处理,但是 RNN 很少用于文本特征提取,其根本原因是它主要以时间序列为目标。此外,由 Ian J. Goodfellow 于 2014 提出的生成对抗性的网络模型,短短两年时间在深度学习生成模型领域取得了显著成果。文中提出了一种新的可用于估计和生成对抗过程模型的框架,并将其作为无监督学习的一种突破。现在它主要用于生成自然图像,但在文本特征提取方面没有取得重大进展。

深度学习中存在一些瓶颈,如监督感知和强化学习都需要大量的数据支持。此外,在推进计划方面表现很差,不稳定的数据质量导致的不可靠、不准确和不公平的问题仍需要改进。由于文本特征提取的固有特性,每种方法都有其优缺点。如果可能的话,可以应用多种提取方法来提取相同的特征。

参考文献:

[1] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[2] QIN S, LU Z. Sparse automatic encoder application in text categorization research[J]. Science Technology and Engineering,2013,13(31):9422-9426.

[3] SINGH V, KUMAR B, PATNAIK T. Feature extraction techniques for handwritten text in various scripts: a survey [J]. International Journal of Soft Computing & Engineering, 2013,3(1):238-241.

[4] SUTO J,ONIGA S,SITAR P P. Feature analysis to human activity recognition[J]. International Journal of Computers Communications & Control,2016,12(1):116-130.

[5] MLADENIC D. Machine learning on non-homogeneous,distributed text data[D]. Ljubljana: University of Ljubljana, 1998.

[6] NIHARIKA S,LATHA V S,LAVANYA D R. A survey on text categorization[J]. International Journal of Computer Trends & Technology,2006,18(3):72-74.

[7] MHASHI M,RADA R,MILI H,et al. Word frequency based indexing and authoring[M]//Computers and writing. [s. l.]:Springer,1992:131-148.

[8] PANINSKI L. Estimation of entropy and mutual information [J]. Neural Computation,2003,15(6):1191-1253.

[9] RUSSAKOFF D B,TOMASI C,ROHLFING T,et al. Image similarity using mutual information of regions[C]//8th European conference on computer vision. Prague,Czech Republic:Springer,2004:596-607.

[10] EVANGELOPOULOS N E. Latent semanticanalysis[J]. Annual Review of Information Science & Technology,2013,4 (6):683-692.

[11] FERCHICHI S E,ZIDI S,LAABIDI K,et al. Feature clustering based MIM for a new feature extraction method[J]. International Journal of Computers Communications & Control, 2013,8(5):699-707.

[12] 罗　燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718-725.

[13] DANUBIANU M,PENTIUC S G,DANUBIANU D M. Data dimensionality reduction for data mining: a combined filter-wrapper framework[J]. International Journal of Computers Communications & Control,2012,7(5):824-831.

[14] ZHOU Yong,LI Youwen,XIA Shixiong. An improved KNN text classification algorithm based on clustering[J]. Journal of Computers,2009,4(3):230-237.

[15] HAN E H,KARYPIS G,KUMAR V. Text categorization using weight adjusted k-nearest neighbor classification[C]// Pacific-Asia conference on knowledge discovery and data mining. Hong Kong:Springer,2001:53-65.

[16] TU A L. A novel probabilistic feature selection method for text classification research[D]. Wuhan: Central China Normal University,2012.

[17] BHATTACHARYA M,DAS A. Genetic algorithm based feature selection in a recognition scheme using adaptive neuro fuzzy techniques[J]. International Journal of Computers Communications & Control,2010,49(8):1421-1422.

[18] BENGIO Y. Learning deep architectures for AI[J]. Foundations & Trends® in Machine Learning,2009,2(1):1-127.