基于深度信念网络的文本分类算法①

陈翠平

(同济大学 计算机软件与理论系, 上海 201804)

摘 要: 随着网络的迅猛发展, 文本分类成为处理和组织大量文档数据的关键技术. 目前已经有许多不同类型的神经网络应用于文本分类, 并且取得良好的效果. 但是, 大部分模型仅采用文档的少量特征作为输入, 没有考虑到足够的信息量; 而当考虑到足够的特征时, 又会发生维数灾难, 导致模型难以训练或者训练时间大幅增加. 利用深度信念网络从文本中抽取特征, 并利用 softmax 回归分类器对抽取后的特征分类. 深度信念网络不仅具有强大的学习能力, 同时还能从高维的原始特征中抽取低维度高度可区分的低维特征, 因此利用深度信念网络来对文本分类, 不仅能够考虑到文档的足够的信息量, 而且能够快速的训练. 并且实验结果也表明利用深度信念网络实现文本分类的性能很好.

关键词: 文本分类; 受限玻尔兹曼机; 深度信念网络; softmax 回归分类器; 文本特征.

Text Categorization Based on Deep Belief Network

CHEN Cui-Ping

(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract: With the rapid development of the network, text categorization has become a key technology in processing and organizing large text. There are already many different types of neural networks applied to text categorization and achieve good results. However, most of the document models use only a small amount of features as input which don't take into account the sufficient amount of information. Considering enough characteristics, the dimensions of the disaster will occur, and resulting in a substantial increase training time, difficulty in training the model. This paper considers using deep belief networks to extract features from the text, then using softmax regression classifier for classification. The deep belief networks not only has a high learning ability, but also can extract highly distinguishable and low-dimensional features from the original high-dimensional features. So taking advantage of the deep belief networks to classify the text, can take into account enough information document amount. And the result shows that the use of deep belief networks achieve good performance for text categorization .

Key words: text categorization; restricted boltzmann machine; deep belief network; softmax regression classifier; feature.

1 引言

随着信息技术的发展, 互联网数据及资源呈现海量特征. 基于内容的信息检索和数据挖掘逐渐成为备受关注的领域. 其中, 文本分类(text categorization, 简称 TC)技术是信息检索和文本挖掘的重要基础, 其主要任务是在预先给定的类别标记(label)集合下, 根据文本内容判定它的类别. 文本分类在自然语言处理与理解、信息组织与管理、内容信息过滤等领域都有着广泛的应用. 文本分类的方法有很多, 典型的有朴素贝叶斯分类器[1]、k 最近邻(kNN, k-Nearest Neighbor)算法[2]、用支持向量机(SVM)[3]建立的分类器和用 BP(Back Propagation)神经网络[4]建立的分类器等, 并且均取得了良好的效果. 并且有学者在这些传统的文本分类器的基础上提出改进的方案[5-9], 使得分类的效果比原来的更加优良. 但是这几种方法均通过浅层神经网络来实现的, 其共同的局限性在于有限样本和计

① 收稿时间:2014-06-17;收到修改稿时间:2014-07-22

算单元情况下对复杂函数的表示能力有限, 针对复杂分类问题其泛化能力受到一定制约[10,11], 且常常面临维数灾难、局部最优及过学习问题等.

深度神经网络通过多层神经网络来训练模型, 不仅能克服这些情况, 同时由于深度神经网络具有多层非线性映射的深层结构, 因此能够利用较少的参数完成复杂的函数逼近, 具有良好的特征学习能力. 此外深度学习还可以通过组合低层特征形成更加抽象的高层表示, 从而能够发现数据的分布式特征表示, 即可通过逐层学习从而获得输入数据的主要驱动变量[12], 且具有强大的从少数样本集中学习数据集本质特征的能力[13,14]. 且深度神经网络可以在训练好之后可以展开成普通的 BP 神经网络, 并利用 BP 神经网络的 BP 算法微调优整个网络, 从而优化整个网络的性能.

2006 年, Hinton 等人提出了深度信念网络(deep belief network, DBN)[15]的逐层贪婪的学习方法[16], 能够避免传统的梯度下降算法针对多隐层训练效果不佳的问题. 目前 DBN 已广泛应用与手写体识别[15]、图像识别[17]以及语音识别[18]等各种领域. 但是应用于文本分类方面的研究仍然比较少. 本文利用 DBN 良好的特征学习能力, 从原始的特征中学习更加抽象且高度可区分的特征, 将得到的特征输入 softmax 回归分类器实现分类. 并且通过实验和 KNN 算法、SVM 分类器以及传统的 BP 神经网络几种常用的文本分类方法进行了比较, 发现基于 DBN 的文本分类器分类的准确率和这几种方法相比, 有一定程度上的提高.

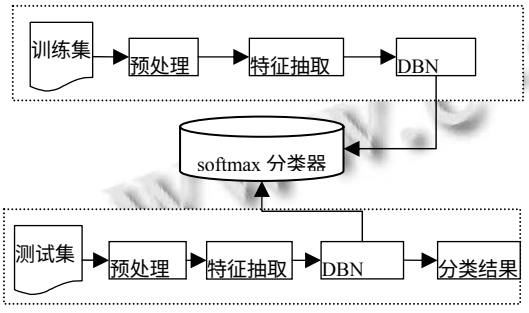

图 1　文本分类模型框架

2　文本分类的过程

由于文本文档是大量字符的集合, 是非结构化或半结构化的数字信息, 不能直接被任何分类器所识别, 必须将其转换成为一个
简洁的、统一的、能够被学习算法和分类器所识别的结构化形式, 才能进行进一步的分析和处理. 如图 1 中文本分类模型框架中描述的一样, 需要对文档进行预处理, 先选择能够代表文档本质特征的元数据, 并运用向量空间模型(Vector Space Model, VSM)的形式存储, 使得计算机能够更加方便的对这些数据做后续的一系列处理, 最终完成分类.

2.1　文档预处理

在本文中, 将所有的文档看成是“词袋”(bag of words)形式, 然后对文档处理, 预处理的过程有以下几步:

①得到文档中所有的正文信息, 即去除文档的标签信息、标点符号、数字, 将所有单词均变成小写形式, 使得文本变成一个只有词条和空格组成的小写的字符串.

② 对文档进行去停止词处理. 为了避免将无用词作为特征, 一般基于停用词表(stop list)去掉“不太相关”的词, 以减少特征总数[19]. 停滞词通常有冠词、助词、代词、介词、连词等. 例如, 尽管“the”、“it”、“and”和“of”等词语在文本中频繁出现, 但是基本上没有区分或预测能力, 因此属于停用词.

③ 利用波特词干分析法处理文档, 得到文档中单词的词根形式. 文档中很多有很多词语的词形发生变化, 而词义却不会有很大区别, 那么只需将词根作为特征[20], 这样就能够在很大程度上减少文本向量的维数. 例如, 名词有单复数不同的形式, 动词有时态的变化, 形容词具有比较级等.

④ 构建单词——数字映射表, 将单词映射成数字, 使得能够方便的将处理后的文档表示成向量空间模型中的特征向量形式, 每一个特征对应一个单词. 例如, 单词 economic 映射成 40, 则在 VSM 表示的文本矩阵中第 40 列表示单词 economic 的权重.

⑤ 统计单词频率, 统计每篇文档中每个单词的个数, 并用 VSM 形式存储.

通过以上几种的预处理工作, 我们就可以把文本集合中的文本表示成特征向量的形式, 使得文本数据结构化, 从而能够便于后续的对文本的处理.

2.2　文档原始特征选择

本文中的原始特征选择的是预处理后文本中单词的 TF-IDF. 其中 TF-IDF 的计算公式如下:

$$TF\text{-}IDF=(TF/Ni)*\lg(N/DF) \quad (1)$$

公式中 TF 是指文档中给定单词的词频,N_i 是文档中单

词的总数; IDF 是逆向文件频率, 是一个单词的重要性的度量, N 表示文档总数, DF 表示包含该单词的文档总数. 由公式(1)可以看出字词的重要性随着它在文件中出现的次数成正比增加, 但同时会随着它在语料库中出现的频率成反比下降, 这样既可以选择在语料库中出现频次比较高的单词, 又可以过滤掉常用词. 因此 TF-IDF 可以用以评估一个单词对于一个文档集或一个语料库中的其中一个文档的重要程度. 选用的 TF-IDF 最高的单词是对语料库最重要的单词而不仅仅是语料库中出现频率最高的单词, 相比而言, 更加能够表示文档的原始特征.

选择单词总 TF-IDF 最高的 3000 的单词作为原始特征, 虽然此时特征维数相对较低, 但是此时的特征是一个个单词的 TF-IDF 的排列组合, 各原始特征之间的各种深层特征仍有待进一步提取. 本文将采用深度信念网络对原始特征进行进一步的提取和降维, 从而抽取低维的高度可分的特征. 最后利用 softmax 回归分类器来对 DBN 训练过后的高度抽象特征进行分类.

3 DBN建模

当文档的原始特征已经被表示成矩阵形式后, 可以通过 DBN 进一步提取更加抽象和高度可分的特征.

3.1 DBN 简介

深度神经网络(DBN)是一种可提取数据深层特征的方法, 它解决了传统多层神经网络难以进行训练的问题. 根据目标函数的不同, 目前已有多种深度网络诞生, 如深度信念网络、Autoencoder、CNN 等. 近年来, 通过 DBN 从语料的底层特征中提出抽象的高层特征, 结合 BP 神经网络, 支持向量机, softmax 回归等分类器对文本进行分类, 已经取得很好的效果. 但是目前国内利用深度信念网络来对文本进行分类的研究依然比较少.

3.2 受限玻尔兹曼机原理

受限玻尔兹曼机(RBM)是一种典型的神经网络, 如图 2 所示. 该网络可视层和隐层中神经元彼此互联(层内无连接), 隐层可获得输入可视层单元的高阶相关性. RBM 中神经元有两种状态: “激活”和“未激活”, 一般用二进制的 1 和 0 表示. 受限玻尔兹曼机一个最主要的优点是在给定可见层节点状态时所有隐藏节点的状态是条件独立于其他隐藏节点的; 反之, 在的情况下, 各个可见层节点的状态亦条件独立. 因给定隐藏节点此, 尽管 RBM 所表示的分布无法有效计算, 但是通过 Gibbs 采样可以得到服从 RBM 所表示分布的随机样本. 此外, Roux 和 Bengio[21]从理论上证明, 只要隐藏层单元的数目足够多, RBM 能够拟合任意离散分布.

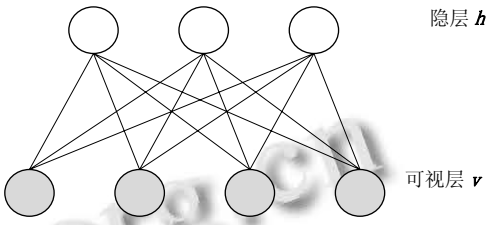

图 2 受限的玻尔兹曼机

3.3 RBM 的能量模型

RBM 是一种基于能量的模型, 如图 2 中所示的 RBM 中, 可见层变量 v 和隐藏层变量 h 的联合配置的能量为:

$$E(v, h \mid \theta) = \frac{1}{2}(v^T W h + b^T v + c^T h) \quad (2)$$

其中 $\boldsymbol{v}$ 代表可视层节点的状态向量, $\boldsymbol{h}$ 代表隐藏层节点的状态向量, $\theta=(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c})$ 为参数集合, $\boldsymbol{W}$ 代表可视层节点与隐藏层节点的连接权值矩阵, $\boldsymbol{b}$、$\boldsymbol{c}$ 分别表示可视层节点和隐藏层节点的偏置向量. 当参数确定后, RBM 的归一化因子(或配分函数)为 $Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}$, 此时基于能量函数, 我们可以得到$(\boldsymbol{v}, \boldsymbol{h})$的联合概率分布:

$$P(v, h \mid \theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)}, \quad (3)$$

隐藏层节点的条件概率如下:

$$p(h_j = 1 \mid v) = \delta(c_j + \sum_i v_i W_{ij}), \quad (4)$$

可视层节点的条件概率如下:

$$p(v_i = 1 \mid h) = \delta(b_i + \sum_j W_{ij} h_j), \quad (5)$$

其中 $\delta(x) = \frac{1}{1+\exp(-x)}$ 为 sigmoid 激活函数.

3.4 DBN 的训练

学习 RBM 的任务是求出参数的值, 来拟合给定的训练数据. 可以通过最大化 RBM 在训练集(假设包含 N 个样本)上的对数似然函数学习得到, 即 $\theta^* = \arg\max L(\theta) = \arg\max \sum_{t=1}^{N} \log P(v^{(t)} \mid \theta)$ 在

计算最优参数 θ 的过程中, 由(3)式可知, 我们首先要计算归一化因子 $Z(\theta)$, 但是 $Z(\theta)$很难计算, 因此由于归一化因子的存在, θ 很难计算出来.

在 2002 年, Hinton[16]提出了 RBM 的一个快速学习算法, 即对比散度(Contrastive Divergence, CD). 首先, 我们将可见单元的状态设置成一个训练样本 v, 利用公式(4)计算所有隐层单元的二值状态. 而在隐层单元的状态 h 确定之后, 我们利用公式(5)来计算在隐藏层确定的情况下, 可视层所有单元的状态, 从而产生可视层的一个重构(reconstruction)v_r. 如果此时 v_r 和 v 一样, 那么得到的隐层 h 就是可视层 v 的另外一种表达, 因此隐含层可以作为可视层输入数据的特征. 同时在训练中可以利用训练样本的状态与计算出来的可视层的状态之间的重构误差来调整 RBM 的参数, 从而使得 v 和 v_r 的重构误差尽可能减小.

RBM 的 CD 算法主要步骤如下:

输入: 一个训练样本 x_0; 隐层单元个数 m; 学习率 ε; 最大训练周期 K.

输出: 连接权重矩阵 W、可见层的偏置向量 $\boldsymbol{a}$、隐层的偏置向量 b.

① 初始化: 令可见层单元的初始状态 $v_1=x_0$; W、a 和 b 为随机的较小数值.

② 根据公式(4)计算隐藏层节点的状态, 即计算 $p(h_{1j}=1 \mid v_1)=\delta(c_j+\sum_i v_{1i}W_{ij})$. 然后令隐藏层节点 h_j 以 $p(h_j=1 \mid v)$ 的概率设置为 1, 否则为 0.

③ 根据第 2 步计算得到的隐藏层状态和公式(5)来重构可视层 v_2. 即计算 $p(v_{2i}=1 \mid h_1)=\delta(b_i+\sum_j W_{ij}h_{1j})$. 并令可视层节点 v_i 以 $p(v_{2i}=1 \mid h_1)$ 的概率设置为 1, 否则为 0.

④ 通过重构后的可视层计算得到的隐藏层节点的概率 $p(h_{2j}=1 \mid v_2)=\delta(c_j+\sum_i v_{2i}W_{ij})$.

⑤ 根据下列公式, 更新 RBM 中的各个参数:

$$W=W+\varepsilon(p(h_{1j}=1 \mid v_1)v_1^T-p(h_2=1 \mid v_2)v_2^T)$$

$$b=b+\varepsilon(v_1-v_2)$$

$$c=c+\varepsilon(p(h_1=1 \mid v_1)-p(h_2=1 \mid v_2))$$

其中, ε 为学习率.

⑥ 取下一个数据样本, 重复 2-5 的过程.

⑦ 2-6 的过程重复迭代 K 次.

DBN 是由一层层的 RBM 堆叠而成的, 在训练 DBN 时, 可以采用贪婪法逐层训练每一层的 RBM. 前一层的 RBM 训练完成后, 将其隐藏层节点的激活概率矢量作为该 RBM 的可视层来训练该层 RBM, 以此类推来训练若干层 RBM, 构建完整的 DBN 网络.

3.5 网络调优

DBN 在训练的过程中是按照逐层贪婪的方法进行训练的, 前一层的 RBM 的误差会逐渐往后层的 RBM 传递但是得不到修正, 而传统神经网络中的 BP 算法可以利用误差反向传播的过程调优整个神经网络. 因此在 DBN 训练之后, 将 DBN 的输出层作为 softmax 回归分类器的输入, 形成一个完整的文本分类神经网络后可以利用 BP 算法调优这个文本分类神经网络.

传统的 BP 神经网络的权值和偏置值均是随机初试化, 再利用 BP 算法调优权值和偏置值, 直到收敛. 而在调优 DBN 的过程中, 可以利用 DBN 的权值初始化 BP 神经网络的各层的权值, 而不是利用随机初始值初始化网络, 将 DBN 展开成 BP 神经网络, 最后利用 BP 算法微调优整个网络的参数, 从而使得网络的分类性能更优.

4 实验结果分析

本文的实验的数据即训练样本的选择采取的是路透社语料库(RCV1-v2 corpus), 该数据集是路透社新的数据集, 是对 RCV1 文档集的一个改进, 修改了其中的一些分类错误. 该数据集共有 80000 篇新闻文章, 包括 103 个类别, 本次试验选取其中 4 类进行实验, 分别为: “C15”、“ECAT”、“GCAT” 和 “MCAT”, 共 9625 篇文章[22], 其中单词总数为 29992. 训练集选择其中 6000 篇(每类 1500 篇)文章, 余下的 3625 篇文章作为测试集. 并且使用常用的准确率来评估分类系统的性能, 准确率的计算公式为:

准确率=分类正确的文档数/文档总数

表 1　分类实验结果

	BP 神经网络	KNN	SVM	DBN
C15	89.08	93.3	95.02	98.28
ECAT	86.88	88.48	94.33	93.62
GCAT	96.07	93.5	97.57	97.43
MCAT	84.53	93.23	95.17	95.43
average	90.01	92.61	95.94	96.33

为了测试基于DBN的网络的分类性能, 我们还在相同的数据集上利用KNN分类器、SVM、传统的BP神经网络分别进行了以下几种分类实验:

① DBN神经网络的节点数分别为3000-1500-750-375-188-100-4共7层, 迭代2000次;

② BP神经网络是3层神经网络, 节点数分别为3000-500-4, 迭代2000次;

③ 使用matlab自带的knnclassify分类器进行knn分类实验;

④ 使用libsvm工具箱, 进行svm实验.

分类过程中, 各分类器均并采用不同的配置参数进行多次实验, 选择其中最优的结果作为实验结果. 实验结果如表1所示 : 由表1可以明显看出, 由DBN深度神经网络训练得到的分类器的分类效果要明显的好于SVM分类器、KNN分类器和BP神经网络分类器等几种浅层网络的分类效果. 因此理论与试验均说明深度神经网络在文本分类方面的分类效果优于浅层神经网络.

5 结语

本文采用了基于深度信念网络抽取低维度高度可区分的特征对文档分类. 并且进行了四种不同类别的实验, 得到DBN在文类分类时能够比其他几种浅层神经网络具有更优的分类效率. 而实验也表明通过DBN抽取文档特征, 并利用 softmax 回归分类器对文档分类这种方法能够得到令人满意的分类结果. 本文的研究中只抽取了单词在文档中的 tf-idf 形式来作为文档的原始特征, 这种原始特征提取的方式过于简单, 如何提取复杂的文档原始特征进行更加快速有效的文档分类是下一步的研究方向.

参考文献

1 McCallum A, Nigam K. A comparison of event models for naive bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. Madison, Wisconsim(32).

2 白莉媛,黄晖,刘素华,阎秋玲.基于自助平均的朴素贝叶斯文本分类器,2007,33(15):190–192.

3 Joachims T. Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning (ECML). Chemnitz, Germany. 1998. 137–142.

4 Ruiz ME, Srinivasan P. Hierarchical neural networks for text categorization. Proc. of SIGIR-99, 22nd ACM International Information Retrieval. 1999(32). 281–282.

5 叶菲,罗景青,俞志富.一种改进的并行处理 SVM 学习算法. 微电子学与计算机,2009,26(2):40–43.

6 Guo GD, Wang H, Bell D, Bi YX, Greer KR. An kNN Model-based Approach and Its Application in Text Categorization. CICLing 2004, LNCS 2945, 2004. 559–570.

7 吴春颖,王士同.一种改进的 KNN Web 文本分类方法.计算机应用研究,2008,25(11):3275–3277.

8 曾砺锋.基于 Rocchio 方法和 k 均值聚类的支持向量机文本分类方法.软件导刊,2008,7(6):37–39.

9 陈世立,高野军.基于神经网络与贝叶斯的混合文本分类研究.情报杂志,2007(5):34–36.

10 Debole F, Scbastiani F. An analysis of the relative hardness of recuters-21578 subsets. Journal of the American Society for Information Science and Technology,2004,56(6): 584–596

11 Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009, 2(1): 1–127.

12 孙志军,薛磊,许阳明,等.深度学习研究综述.计算机应用研究,2012,29(8):2806–2810.

13 Bengio Y, Delalleau O. On the expressive power of deep architectures. Proc of the 14th International Conference on Discovery Science. Berlin: Springer-Verlag, 2011: 18–36.

14 Bengio Y, Lecun Y. Scaling learning algorithms towards AI. In: Bottou L, Chapelle O, Decosted, et al. eds. Large-Scale Kernel Machines. Cambridge: MIT Press, 2007: 321–358.

15 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507.

16 Hinton GE, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554.

17 Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proc. of the 26th Annual International Conference on Machine Learning. ACM. 2009. 609–616.

18 Dong Y, Li D. Deep learning and its applications to signal and information processing. IEEE Signal Process Mag,

2011, 28: 145–154.

19 Fox C. Lexical analysis and stoplists. Information Retrieval: Data Structure and Algorithms, 1992, 7: 102–130.

20 Frakes WB. Stemming algorithms. Information Retrieval: Data Structure and Algorithms, 1992, 3: 131–160.

21 Roux NL, Bengio Y. Representational power of restricted Boltzmann machines and deep belief networks. Neural Computation, 2006, 20(6): 1631–1649.

22 Cai D, He XF. Manifold adaptive experimental design for text categorization. IEEE Trans. on Knowledge and Data Engineering, 2011.