情报工程 ISSN 2095-915X
第2卷 第5期 30-40 2016 年 10 月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.5 30-40 Oct 2016

doi:10.3772/j.issn.2095-915x.2016.05.004

基于混合深度信念网络的多类文本表示与分类方法

翟文洁[1]，闫琰[2]，张博文[1]，殷绪成[1]

（1. 北京科技大学计算机科学与技术系 北京 100083；2. 中国矿业大学计算机科学与技术系 北京 100083）

摘要：本文开展了基于混合深度信念网络的多类文本表示与分类方法的研究，以解决传统的 Bag-of-Words（BOW）表示方法忽略文本语义信息、特征提取存在高维度高稀疏的问题。文章基于文本关键字，针对多类的分类任务（如新闻文本和生物医学文本），以关键字的词向量表示作为文本输入，同时结合深度信念网络（Deep Belief Network，DBN）和深度玻尔兹曼机网络（Deep Boltzmann Machine，DBM），设计了一种混合深度信念网络（Hybrid Deep Belief Network，HDBN）模型。文本分类和文本检索的实验结果表明，基于词向量嵌入的深度学习模型在性能上优于传统方法。此外，通过二维空间可视化实验，由 HDBN 模型提取的高层文本表示具有高内聚低耦合的特点。

关键词：文本分类，文本表示，深度学习，深度信念网络

中图分类号：TP391

A Model for Text Representation and Classification Based on Hybrid Deep Belief Networks

ZHAI WenJie[1], YAN Yan[2], ZHANG BoWen[1], YIN XuCheng[1]

(1. Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing 10083, China;
2. Department of Computer Science and Technology, China University of Mining and Technology, Beijing 10083, China)

基金项目：本文受国家自然科学基金项目：结合前馈和反馈机制的自然场景文本识别技术（61473036）资助。

作者简介：翟文洁（1991–），硕士生，研究方向：信息检索；闫琰（1992–），博士，研究方向：自然语言处理；张博文（1992–），博士生，研究方向：信息检索，机器学习；殷绪成（1977–），通迅作者，博士，教授，博导，北京科技大学计算机与通信工程学院计算机科学与技术系模式识别技术创新实验室主任，IEEE Senior Member，研究方向：文档分析与识别、模式识别信息检索与推荐系统。

Abstract: This paper developed a model for text representation and classification based on hybrid deep belief networks, in order to solve the problem of traditional text representation methcod (Bag-of-Words), which ignores the semantic relations and whose feature extraction is high-dimensional and high-sparse. Based on the text keywords, we explored the word vector of keywords as the input for multiple classification tasks, such as news and biomedicine texts, and we also proposed a new model —HDBN (Hybrid Deep Belief Network) which is based on the integration of DBN (Deep Belief Network) and DBM (Deep Boltzmann Machine). The results of text categorization and text retrieval showed that the HDBN model can performed better than the traditional methods. Moreover, the results of two-dimensional spatial visualization also indicated that high-level text representation based on the HDBN model presented the character of high cohesion and low coupling.

Keywords: Text classification, text representation, deep learning, deep belief networks

1 研究背景

在“信息过载”时代，如何有效地管理、过滤和筛选信息成为一项重要的研究课题。在网络信息中，文本作为主要的信息承载途径占据着重要地位。为方便用户准确地定位所需的信息，使用文本分类技术可以在较大程度上处理和解决信息杂乱的问题。为了完成文本分类，首先要将文本转化为可处理的形式。因此，文本表示作为文本分类的基石，它的准确度几乎决定了自然语言处理的效果。当前文本表示方法多基于词袋模型（Bag-of-Words，BOW）或向量空间模型（Vector Space Model，VSM），默认单词彼此独立，忽略了句子上下文关系。然而随着文本类型多样、主题细化，这种浅层文本表示缺失文档的语义信息，难以处理目前复杂的分类问题。

深度学习的发展给文本分类带来新的机遇。它作为当前解决大数据难题的重要学习理论，对文本表示与文本分类具有一定的借鉴作用。目前深度学习已经成功应用于多种模式分类问题，数据压缩、目标检测和跟踪、信息检索、机器翻译和语音识别等，在自然语言处理领域方面也取得了一定的成果。使用基于深度学习的方法，可以更好得挖掘蕴含在文本中的复杂语义关系，与具体任务紧密结合。同时，伴随互联网规模的扩张和多媒体的发展，大规模训练数据以及机器设备性能的重大提升，高性能 GPU 和 CPU 集群提供了强大的计算能力，给深度学习提供了一个技术革新的舞台。

2 研究现状分析

本文主要研究文本表示和分类方法。文章以文本为研究对象，为实现复杂海量文本的准确分类，采用基于混合深度信念网络的方法，提升分类的效果。

2.1 文本表示的现状

由非结构化或半结构化的数字信息组成的文本文档不能直接被分类器所识别，须转换为统一的、可被学习算法和分类器所识别的结构化形式[1]。可用图或向量来表示。随着文本的多样化发展和研究的深入，文本表示模型开始从特征选择转变为特征提取，即基于语义的文本表示方法。

基于语义的文本表示方法主要有线性和分布式两种。线性表示方法主要有 LDA（Latent

Dirichlet Allocation）[2]模型、LSI（Latent Semantic Indexing）[3]、PLSI（Probabilistic Latent Semantic Indexing）[4]，监督的LDA以及SSI（Supervised Semantic Indexing）等。这些方法提取得到的表示可以认为是文档的深层表示，特征提取的方式在一定程度上考虑了文本的语义特征。此外，基于文档输入表示（Word Embedding）的分布式表示借鉴了自然语言处理技术，赋予不同单词不同的含义，目前已经验证了理论上的合理性。

2.2 文本分类的现状

文本分类的方法主要有两类：基于规则的分类方法和基于统计的分类方法。其中，基于规则的分类方法多需要该领域的知识、规则库作支撑，具有针对性强但可扩展性不足的特点。基于统计的学习方法通过对样本统计和计算，建立相应的数据模型学习参数并完成分类器的训练。

目前，大量的基于统计的机器学习方法被应用于文本分类系统中，应用最早的机器学习方法是朴素贝叶斯（Naive Bayes，NB）。随后，几乎所有重要的机器学习算法在文本分类领域都得到了应用，比如K近邻算法（K Nearest Neighbor，KNN）、支持向量机（Support Vector Machine，SVM）、神经网络（Neural Net,Work,NN）、最小二乘和决策树等。其中，向量空间模型的应用是文本分类近几年来最重要的进展之一。

2.3 深度学习的现状

深度学习起源于人工神经网络的研究，是基于深度神经网络（Deep Neural Network，DNN）一类学习方法的统称。DNN源于出现于上世纪四十年代的传统神经网络，其主要目的是尝试通过模拟人脑认知的机理来解决各种机器问题。

机器学习的第二次浪潮开始于2006年Hinton等人发表在Science上的一篇深度学习的文章。他提出了一种深度信念网络（Deep Belief Network，DBN），即基于层叠的RBM（Restricted Boltzmann Machine）学习算法，通过无监督的贪心逐层预训练，配合有监督的微调训练方法，解决了深度学习模型优化困难的问题，使得以往制约DNN发展的各种限制被一一解决，DNN开始了快速的发展。

在此，介绍本文使用到的两种深度学习模型。深度信念网络（DBN）和深度玻尔兹曼机（Deep Boltzmann Machine，DBM）[5]，二者在网络节点选取和网络结构上基本一致，都采取受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）模型作为网络的基本建模单元。

深度学习模型具有复杂的多层结构，通常都是基于RBM网络实现相邻两层之间连接。RBM网络共有2层（如图1），其中第一层为可视层（v）也叫输入层，它由m个可视节点组成；第二层为隐含层（h），也就是特征提取层，由n个隐藏节点组成。在已知v的情况下，所有的隐藏节点之间是条件独立的，即$P(h/v)=P(h_1/v)...P(h_n/v)$。同理，在已知隐藏层$h$的情况下，所有的可视节点也都是条件独立的，即层内节点无连接，层间节点全连接（如图1）。

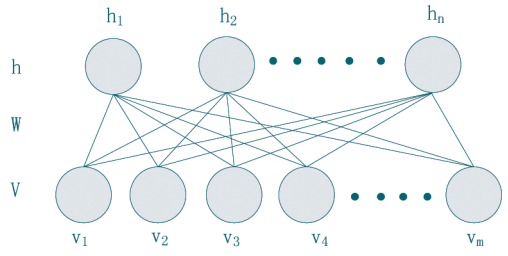

图1 受限玻尔兹曼机模型示意图

在采样步数足够大的情况下，基于RBM模型的对称结构，以及模型中各个神经元状态条件独立，可以得到服从RBM所定义的样本分布，

也就是隐藏层表示 h。

通过 RBM 模型训练得到层间连接的权重矩阵和偏移量，提供 BP 神经网络进行初始化训练，还可以对数据进行编码，然后用监督学习方法进行分类或者回归应用。增加隐藏层的层数，可得到深度玻尔兹曼机（DBM）；如果在靠近可视层的部分使用贝叶斯信念网络（即有向图模型），而在最远离可视层的部分使用 RBM，可得到深度信念网络（DBN）。

DBN 是由多个限制玻尔兹曼机（RBMs）层组成的一个神经网络，既可以被看作一个生成模型，也可以当作判别模型，这些网络被限制为一个可视层和一个隐层，层间存在连接，但层内的单元间不存在连接 [6]。DBM 和 DBN 类似，都是由 RBMs 组成的网络结构，但是 DBM 是无向图模型，DBN 是有向的图模型结构。DBN 相比于普通神经网络，模型训练时间会显著的减少。

2.4 现状分析

文本分类和检索任务中至关重要的一步就是文本表示，它决定了语义索引的正确率。对于浅层文本表示（特征选择），存在语义的缺失的问题；对于基于线性计算的模型的深层文本表示，分类器训练时加入了阈值的选择，破坏了文本自我学习的能力。针对多标签和多类问题的文本分类，存在忽略标签依赖关系，泛化能力不足的问题。

针对以上问题，学术界开展了基于深度学习的文本表示和分类的研究。例如，Hinton 和 Salakhutdinov 提出了两层的 RSM（Replicated Softmax Model）模型 [9]，实验结果表明其效果优于 LDA 方法。但是该模型基于权重共享而且仅有两层，在降维过程中文档缺失的信息比较多，不能足够充分的学习文档的表示，导致模型最后学出的不同文档表示的差异性不大。因此，本文提出具有两层的 DBN 模型和两层的 DBM 模型的 HDBN 模型。

3 HDBN（Hybrid Deep Boltzmann Network）设计

3.1 模型设计

针对传统表示方法高维度高稀疏和缺失语义的缺点，结合深度学习有效提取特征的优势，本文提出了一种基于 DBN 融合 DBM 改进的模型，我们称为 HDBN（Hybrid Deep Belief Network）。

首先在模型的底层用 DBM 进行初降维，使得最大程度地保留文档的信息；然后结合 DBN 进行再次降维，使得更好得提取高层文本特征。HDBN 模型依然遵循标准的 DBN“无监督训练 + 有监督微调”的模型训练方式，通过引入 DBM 初降维，HDBN 模型能更加准确地学习到文档的向量表示。

本文选择两层的 DBN 模型和两层的 DBM 模型（如图 2）。

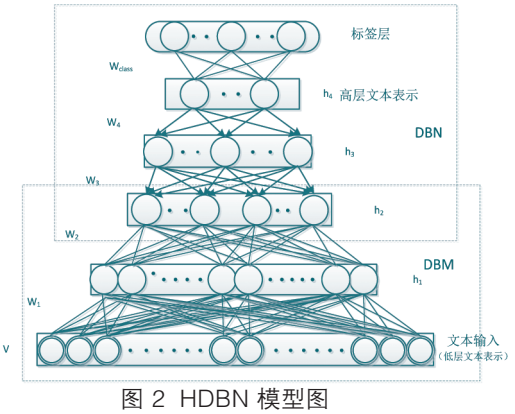

图 2 HDBN 模型图

3.1.1 原理分析

DBM 是由两层 RBM 组成的无向图连接模型，每层节点采样值均由上下两层连接节点共同计算（如图 3）。其中 $q(h_2/v)$ 是后验概率分布，

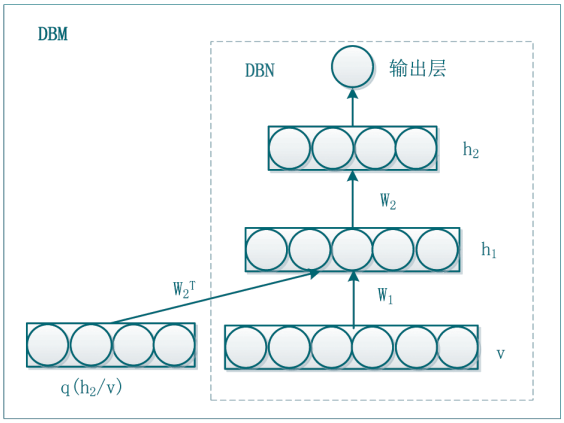

图 3 DBM 和 DBN 模型构造图

代表隐层的重构表示。DBN 是由两层 RBM 组成的有向图连接模型，在预训练过程中，下面层是输入，上面层是输出。当所有层训练结束后，由最上层开始向下有监督的进行微调。

本模型选用两层的 DBM 有两个原因。

其一，实验表明：当 DBM 层数超过两层以后模型的效果不如多层的 DBN，虽然 DBN 在开始训练时容易产生过拟合现象，但是在高层时可以保持很好的性能。其二，DBM 模型训练的复杂度以及训练时间是相同层数的 DBN 模型的三倍 [7]。

在 HDBN 模型中，首先选取两层 DBM 对文本初降维，然后基于 DBN 模型继续训练，这样既能保证提取文本特征又可以减少训练的时间，降低训练的复杂度。

3.1.2　框架分析

HDBN 模型一共有四层（如图 2）。v 是可视层同时也是模型的输入层，每一篇文档用一个固定维度（长度）的向量表示，标签层是模型的标签预测输出。h_1，h_2，h_3 和 h_4 分别对应四层隐层，W_1，W_2，W_3 和 W_4 分别对应层间的连接权重。图中，每一层的节点个数都是不同的，相同层节点之间不连接，不同层节点之间全连接。v 和 h_1 以及 h_1 和 h_2 这两层之间是无向图全连接（DBM 模型），h_2 和以及和之间是有向图全连接（DBN 模型）。模型输入是一个固定长度的向量，先由 h_1 和 h_2 这两层组成的 DBM 训练，h_2 是 DBM 的输出层，同时也是由 h_3 以及 h_4 构成两层的 DBN 模型的输入。h_4 是 HDBN 模型的输出层，代表当前文档的特征表示。对比可视层（v），这一层为高层文本表示，文本分类和检索任务都是基于这个向量计算。

能量模型的目的是捕获变量之间的相关性，因此在训练模型参数时，能否把最优解问题嵌入到能量函数中求解至关重要。为了更好的优化模型的参数，Hinton 等人引入了能量 [8]。其中，RBM 的能量函数定义为：

$$E(v,h) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i \quad (3-1)$$

公式（3-1）代表每个可视节点和隐藏节点连接结构的能量函数。其中，n 是隐层节点的数目，m 是可视层节点的数目，b 和 c 分别是可视层和隐藏层的偏量。RBM 模型的目标函数是累加所有可视节点和隐藏节点取值的能量，因此，对于每个样本都需要统计它对应的所有隐藏节点的取值，这样才可以计算总能量，而这一计算也将面临指数级别的复杂度。一种有效的解决方法是把求解能量模型的问题转换到求解概率的问题。可视节点和隐藏节点的联合概率为：

$$P(v,h) = \frac{e^{-E(v,h)}}{\sum_{v,h} e^{-E(v,h)}} \quad (3-2)$$

通过引入这个概率可以简化求解能量函数，求解的目标是使得能量值最小。在统计力学上有这样一个理论，能量低的状态要比能量高的状态发生的概率高，因此我们要最大化这个概率。引入自由能量函数：

$$FreeEnergy(v) = -ln \sum_{h} e^{-E(v,h)} \quad (3-3)$$

因此

$$P(v)=\frac{e^{-FreeEnergy(v)}}{Z},\ Z=\sum_{v,h}e^{-E(v,h)} \quad (3-4)$$

其中 Z 为归一化因子。于是联合概率 *P(v)* 可以转化为：

$$lnp(v)=-FreeEnergy(v)-lnZ \quad (3-5)$$

公式（3-5）中右边第一项是整个网络的自由能量函数总和的负值，左边是似然函数。

3.2 语义特征表示

本节将进一步探索基于 HDBN 模型的文本输入表示方法，以达到更好地提取特征的目标。

（1）基于词向量嵌入的高维度特征表示

在 BOW 表示形式中，每一个元素代表着当前单词在当前文本中出现的次数，使用一个固定长度（单词个数 *1）的行向量。在基于词向量嵌入的表示中，我们用每一个单词对应的词向量（50 维）去替代 BOW 中相应的单词（1 维），单词个数 *50（见图 4）。用加权系数表示这个单词在当前文本中的重要程度。原始的 BOW 是现在变为一个的向量。

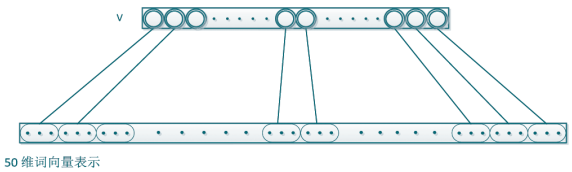

图 4 嵌入词向量的语义特征表示

（2）基于词向量嵌入的关键词特征表示

由于 BOW 表示本身就比较稀疏，导致（1）中的高维度特征向量表示维度更大，使得模型在训练时面临着维度灾难（单词个数 *50）。因此，我们还尝试了基于关键词的词向量嵌入表示形式。这不仅降低了文本输入向量的维度，而且还大大提高了训练的效率。在实验中，我们基于 *TF-IDF* 选取文档的关键词，与普通的 *TF-IDF* 不同的是，我们引入文档的标签，设计标签的权重计算如公式（3-6）：

$$IDF=log\frac{N\times n}{m+k} \quad (3-6)$$

其中 *N* 是文档的个数，*n* 是属于当前类别且包含当前单词的文档个数，*m* 是属于当前类别的文档个数，*k* 表示不属于当前类别但包含当前单词的文档个数。由公式（3-6）计算文档中每个单词的 TF-IDF，然后基于这些数值排序，选取前 40 个单词做为关键词。对比于高维的词向量表示，这个称为基于关键词的低维度词向量表示。

4 实验

实验分三部分。第一部分实验分析影响 RBM 模型（深度模型的基本组成部分）性能的主要参数，以便 HDBN 模型在预训练阶段设置更好的参数。第二部分实验分析验证 HDBN 模型的有效性。为了和已有模型（RSM）做对比，这个模型的输入基于 BOW 特征表示，因此在第一、二部分实验都采用 BOW 的表示方法。在验证模型有效性基础之上，第三部分探索新的文档输入表示对模型的影响。

4.1 数据集描述

我们选取一个新闻数据集（BBC News Data[①]）和一个生物医学数据集（Scientific Medical Abstract[②]）。BBC news data 包括 2,225 篇文档，分别对应商业、娱乐、政治、运动和科技 5 个主题。随机选取 1569 篇文档作为训练集，656 篇文档作为测试集。Scientific Medical Abstract 包括 39 类，平均每类样本文档数目为 300 篇。

① http://www.bbc.co.uk/news/business/market_data/overview/
② http://tinyurl.com/m2c8se6

4.2　实验对比方法

我们采用 DBN、RSM+ 和 DNM 模型进行对比。此外，对于生物医学数据集还与 MTI、MeshUP 和 CNN 三种方法进行了对比。

（1）RBM 网络是构成深度学习模型的主要基本结构，因此在实验部分，我们首先实验并分析影响 RBM 性能的主要参数，然后再对比分析 DBN（由四层 RBM 组成）和 HDBN 模型。

（2）RSM（Replicated Softmax Model）　是 Hinton 和 Salakhutdinov 基于 RBM 设计的一个自动提取低维语义表示的网络，由可视层和隐层组成的一种无向图模型，如图 5。

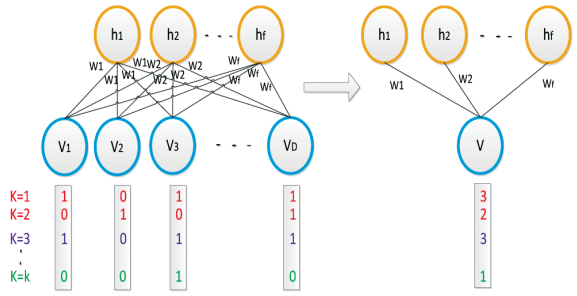

图 5　RSM 模型图

图 5 中蓝色的圈圈表示可视层节点，上图中一共有 D 个可视节点，数值 D 也代表当前文本的单词数目；黄色的圈圈代表隐藏层节点，一共有 f 个隐层节点。左边每一列向量对应每一个单词的表示，右边是对左边所有节点采样的统计结果。当文本采样 D 次后（D 是变量，表示每篇文档的长度），V 就是当前整个文本的表示，也即可视层的输入。

（3）DNM（Deep Network Machines）：　与 HDBN 模型不同的是，我们把 DBM 做高层输出，DBN 模型作低层输入，并称之为 DNM，如图 6。

（4）　MTI[10]（Medical Text Indexer，MTI ③）是一种基于医学主题词表 MeSH 和

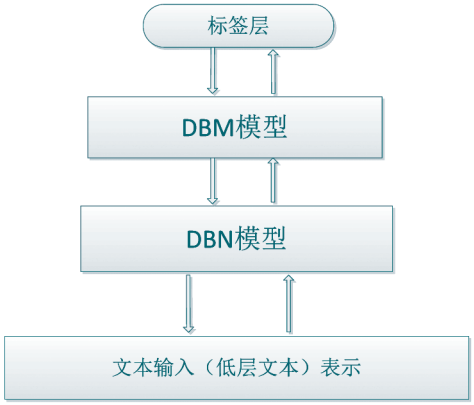

图 6　DNM 模型图

MEDLINE 数据库的自动文献索引系统。MTI 的目的是从 UMLS 概念以及 PubMed 相关文献两条不同路径推荐 MeSH 主题词，主要由两部分组成：MetaMap 映射和 PubMed 相关文献查找。MetaMap 是一种将生物医学文本映射到 UMLS 超级叙词表，或者从文本中挖掘叙词表中概念的程序。它对由这两部分得到的结果，通过聚类或者合并的后处理调整权重，得到 MeSH 的检索排序列表。

（5）MeshUP[11] 是一种结合不同的机器学习方法提升检索性能的系统，具有很高的可扩展能力，基于 MeSH 检索列表结果能显著提升生物医学领域的信息检索性能。

（6）CNN（卷积神经网络）[12] 可以看成是 BP 网络的一个扩展，由多层神经元有规律的彼此连接构成，同时融合了局部感受野、共享权值和空间域或时间域上的次采样这 3 种结构性的方法。局部感受野是指网络中每一层的神经元只与上一层局部领域内的神经单元连接，提取初级的视觉特征；共享权值是指同一张特征图中的神经元共用相同的权值，这样可以减少卷积神经网络的参数，局部感受野和权值共享使得卷积神经网络具有平移不变性；次采样可以减少特征图的分辨率，从而减少对位移、缩放和扭曲的敏感度。

③ http://ii.nlm.nih.gov/MTI/index.shtml

4.3 实验设置

本文对这两个数据集统一做去除停用词、词干化处理，按照词频排序，选取频率在前 2000 的单词作为 BOW 的特征并统计出现频率。为了提高模型的训练效率，把这些样本分为多个子集分组训练，每个子集包含 100 个样本。

对比本文提出的 HDBN 模型（两层 DBN 和两层 DBM），实验中 DNM 和 DBN 模型也都设置为四层。统一设置预训练阶段迭代步数为 50 次，模型更新参数为 0.01。并设置三组不同的学习率（Learning_Rate）：0.01、0.001 和 0.0001；最高层节点个数：50，100，128，500 和 1000。

在实验对比中，直接引用了 Mikolov[13] 已公开的词向量词典。用每个单词对应的词向量表示 BOW 中的每一维，因此高维度的特征向量表示为 100,000 维（2000*50），基于关键词的特征表示为 2000 维（40*50）。此外，在 BOW 特征表示中，每一维度的数值代表着当前词在文档中出现的次数，在词向量嵌入的扩展表示中，单词的权重基于词频和 TF−IDF 的加权值。

5 实验结果和分析

首先实验并分析参数对基本模型（RBM）的影响；然后对比 RSM+，DBN，DNM 和 HDBN 模型，以及 MTI、MeshUP、CNN 模型，并可视化分析模型优缺点；最后尝试分析基于词向量嵌入的表示方式对模型的影响。

5.1 不同模型的对比分析

表 1 是不同模型在 BBC 数据集上文本分类和文本检索任务的正确率。

对比 HDBN 模型和 RSM+ 模型，从表 1 数值分析得到：HDBN 能够检索到更多相似的文档。对同一个待测样本基于两种模型检索出来的相似文档相同。相同样本用 HDBN 模型计算得到的余弦值越小，意味着测试文档在这个模型的低维空间中表示的越细化。因此 HDBN 模型能够更加深层地学习到样本的不同之处，这能够帮助新的测试文档从训练文档中检索到更加相似的文档。

对比 HDBN 和 DBN 模型得到：在两个模型的层数相同（都是四层）情况下，HDBN 先用 DBM 模型训练提取文本特征比 DBN 要更精确。这是因为，在 HDBN 模型中前两层是基于 DBM 预训练，每一个隐层节点的状态都由和它直接相连的上下层节点共同计算决定的，DBM 模型是无向图模型，能得到更好的降维表示。因此 HDBN 模型可以更灵活调整参数，使得上层的 DBN 在训练时基于 DBM 的输出更好地提取文本特征。我们还发现，用 DBM 作为初始特征提取，还可以有效降低模型噪音。

对比 HDBN 和 DNM 模型得到： HDBN 的效果优于 DNM。同时对比 DBN 和 DNM 模型，前两层模型都是基于 RBM 提取特征，所以第二层的输出是一样的，也就是后面两层中 DNM 的性能要优于 DBN。这是因为基于 DBM 引入先验知识，无向图计算隐层节点状态值，更好得提取高层文本表示。

我们对前两层得到的文档表示向量进行可视化实验。发现对于同样的文档，HDBN 得到的文档向量表示在计算相似度的时候余弦距离更小一些。这就表明基于 DBM 模块训练得到的文档之间差异性更大，所以在训练分类器时更容易区分特征，也就更容易能把样本分对。

对比层次分类和 HDBN 模型分类结果（表 2）：HDBN 模型是基于 flat 方法分类，所以在多标签数据集上的分类效果不如层次分类方法。从实验结果来看，HDBN 模型的特征提取要优于只基于 CNN 模型提取的特征，但是不如 MTI 和 MU。即使我们一直在优化模型，但是 Medical 数据集的分类性能仍然不是很高。实验发现，由于

生物医学摘要文本只含有文章题目和摘要信息，大量的专业词和缩写词，而且每篇文档的标签数目不确定，BOW 表示会面临更大的稀疏性，模型的连接权重和节点个数也会增加模型训练的复杂度。此外，数据集自带的错误信息也是导致分类效果不佳的一个原因。

表 1 不同模型的文本分类和检索性能（%）

数据集	模型	学习率	文本分类					文本检索				
			输出层节点个数					输出层节点个数				
			50	100	128	512	1000	50	100	128	512	1000
BBC	RSM+	0.01	94.04	94.65	95.41	95.11	95.87	93.44	93.29	92.98	93.90	94.81
		0.001	95.11	96.48	96.64	94.65	96.64	94.66	95.34	96.18	95.79	96.03
		0.0001	60.40	94.65	94.80	96.79	94.95	70.88	91.76	94.05	95.88	95.42
	DBN	0.01	95.31	94.81	96.03	95.06	95.05	93.18	94.10	94.30	94.13	94.69
	DNM	0.01	95.49	94.89	96.01	94.94	95.08	92.96	94.28	94.68	94.51	95.87
	HDBN	**0.01**	**97.40**	**97.40**	**98.01**	**97.09**	**97.25**	**95.42**	**96.02**	**96.94**	**96.48**	**96.94**

表 2 不同的特征表示在 medical 数据集上的性能对比（%）

特征 / 模型	BOW+				DSE			
	MiP	MiR	MiF	MiS	MiP	MiR	MiF	MiS
MTI	62.03	63.86	62.93	57.44	52.27	52.11	52.19	51.12
MU	63.95	59.62	61.71	56.68	50.66	50.20	50.43	50.22
CNN	53.74	51.03	52.35	51.22	55.42	53.82	54.61	52.42
HDBN	57.39	54.50	55.91	53.16	59.22	57.20	58.19	54.47

5.2 不同的输入特征对模型的影响和分析

这部分实验对比两种基于词向量嵌入表示的方法。第一个是基于词向量嵌入的高维度特征表示，实验结果不是很理想。观察实验结果发现，对于每一个样本，模型预测的类别都是一样的。进一步可视化每层的特征表示发现，在误差反向回传的时候，即使增加迭代步数，精度依然没有改变。此外，对于不同的样本，每层的神经元节点表示是一样的。可视化分析相连两层节点之间不同的连接权值，最后得到的加权结果却是一样的。也就是说，对于不同的样本，在训过程中逐渐减少了样本的差异性，使得最后把它看成同一个样本处理，导致模型在最高层得到的向量表示一样，因此经过 Softmax 函数预测出的文档标签也是相同的。

造成该现象的原因有两个：一是 HDBN 模型的层数太少，导致隐藏层之间没有充分的连接和降维。简单的四层可能不能满足输入长度有 100,000 维度的节点。另外 100,000 维度的输入过于稀疏，高层特征提取时忽略了文本之间的差异，导致节点的表示相同。我们统计每一篇文档处理之后的平均长度，发现文档仅有 100 个单词，即使使用 50 维的词向量去代替 BOW 的表示，非零元素的个数也仅有 5,000 维，仍然有 95,000 的元素都是 0。

针对以上两个原因，我们分别提出了解决方案。一是尝试增加模型的层数。把模型从 4 层增加到 8 层和 10 层，实验结果还是每层节点的状态值相同，但是层间连接的 RBM 的权值是不同的。针对第二个稀疏表示的问题，可以对输入向量进行归一化处理。但是这种处理也没有提高精度，因为虽然这些 0 元素都变成了非零元素，但是在归一化之后被另外一个固定的数值替换。

用词向量嵌入的高维度向量表示效果不是很好，但是我们仍然认为这个带有语义表示的词向量会对模型有一定的促进作用，因此我们尝试降低输入维度，也就是第二种基于关键词提取的词向量嵌入表示。这种特征表示方法首先是提取每篇文档的关键词，然后再用词向量去代替这些关键词。

表 3 是基于关键词的词向量嵌入特征表示实验结果。对比表 3（基于词向量表示）和表 1（基于 BOW 表示）中可以观察得到，基于词向量表示对模型提取高层特征有促进作用。基于词向量和 TF-IDF 表示比 BOW 表示获得更高的精度是因为前者包含两层语义。从文档的输入来看，同一类别不同文档的表示，用基于 TF-IDF 选关键字的表示比用 BOW 表示有更大的差异，这是第一层语义。基于这种方法选择的关键词，文档表示能过滤掉在相同类别中出现的高频词汇和保留相异类别中出现的低频词汇。这种更具差异性的文档输入能够帮助待测样本检索到更加相似和相关的文档。第二层语义是 50 维的词向量表示。这种词向量的表示使得映射在相同维度空间中，相似的文档更加接近。基于相同的模型（HDBN），可视化实验结果得到：以 TF-IDF 输入训练得到的文档表示比 BOW 输入训练得到的文档表示在二维的空间里面更加集中图 7。图 7 是可视化结果，图中相同标签的样本用同一种颜色表示。

表 3 基于关键词的词向量表示在 HDBN 模型上的分类和检索性能（%）

数据集	文本分类					文本检索				
	输出层节点个数					输出层节点个数				
	50	100	128	512	1000	50	100	128	512	1000
BBC data	98.41	98.84	99.35	98.82	97.76	97.26	96.97	98.05	97.52	98.58

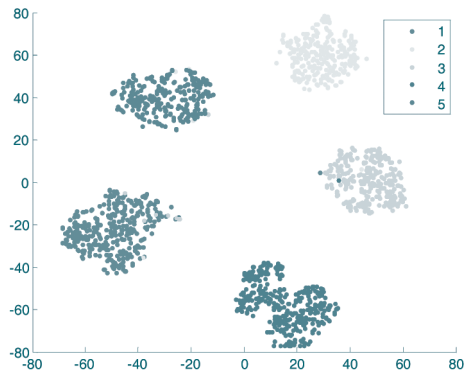

图 7 BBC 数据集上的文本表示可视化

6 小结

针对简单的多类单标签文本分类，本文提出了一种基于混合深度信念网络（HDBN）的文本表示方法。HDBN 模型由无监督预训练和有监督微调两部分组成。首先基于 DBM 无向图采样隐藏层节点状态值，通过引入先验知识，既能有效提取文本特征进行文本初降维，又能去除文本输入带来的噪声；然后基于 DBN 提取高层文本表示。在新闻数据集上的实验结果表明，HDBN 模型要优于别的方法，在生物医学数据集上的实验结果表明，HDBN 优于 CNN 但是不如 MTI 和 MeshUP。除此之外，还检测了不同的文本输入表示以及 RBM 模型节点参数的设置对 HDBN 模型性能的影响，并做可视化分析。我们用已经训练好的文档输入表示代替 BOW 特征表示，实验结果表明基于关键词且嵌入词向量的文本输入表示更利于模型提取高层文本表示。

参考文献

[1] Debole F, Sebastiani F. Supervised Term Weighting for Automated Text Categorization[M]. Text Mining and Its Applications. Springer Berlin Heidelberg, 2004: 81-97.

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of machine Learning research, 2003, 3: 993-1022.

[3] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by Latent Semantic Analysis[J]. JAsIs, 1990, 41(6): 391-407.

[4] Hofmann T. Probabilistic Latent Semantic Analysis[J]. Proc of Uncertainty in Artificial Intelligence Uai', 2010, 25(4):289—296.

[5] Hinton G E, Salakhutdinov R R. A Better Way to Pretrain Deep Boltzmann Machines[J]. Advances in Neural Information Processing Systems, 2012: 2447-2455.

[6] Kim J, Nam J, Gurevych I. Learning Semantics with Deep Belief Network for Cross-Language Information Retrieval[C]// International Conference on Computational Linguistics. 2012:579-588.

[7] Salakhutdinov R, Hinton G E. Replicated Softmax: an Undirected Topic Model[C]// Advances in Neural Information Processing Systems 22:, Conference on Neural Information Processing Systems 2009. Proceedings of A Meeting Held 7-10 December 2009, Vancouver, British Columbia, Canada. 2009:1607-1614.

[8] Salakhutdinov R, Larochelle H. Efficient Learning of Deep Boltzmann Machines[J]. 2010, 9(8):693-700.

[9] Hinton G E, Osindero S, Teh Y W. A fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[10] Kim G R, Aronson A R, Mork J G, et al. Application of a Medical Text Indexer to an Online Dermatology Atlas[J]. Studies in Health Technology & Informatics, 2004, 107(Pt 1):287-91.

[11] Trieschnigg D, Pezik P, Lee V, et al. MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval[J]. Bioinformatics, 2009, 25(11): 1412-1418.

[12] Collobert R, Weston J, Bottou L, et al. Natural Language Processing (Almost) from Scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[13] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.