

基于方差权重因子选词的 SIF 句向量模型

孙 毅, 裘杭萍, 康睿智

(中国人民解放军陆军工程大学 指挥控制工程学院, 南京 210000)

摘 要: 针对平滑反频率(SIF)模型在文本分类和情感分析中性能较差的问题,在 SIF 模型的基础上,根据单词在不同分类任务类别中的分布情况,计算其对任务贡献度的方差权重(VW)因子,建立一种 VW 因子选词句向量模型 CwVW-SIF。在标准文本分类数据集和情感分析数据集上进行测试,结果表明,CwVW-SIF 相对 SIF 模型具有较高的分类精度。

关键词: 平滑反频率;句向量;方差权重;文本分类;情感分析

开放科学(资源服务)标志码(OSID):



中文引用格式: 孙毅,裘杭萍,康睿智. 基于方差权重因子选词的 SIF 句向量模型[J]. 计算机工程,2019,45(9):204-210,234.

英文引用格式: SUN Yi, QIU Hangping, KANG Ruizhi. SIF sentence vector model based on word selection by variance weight factor[J]. Computer Engineering, 2019, 45(9): 204-210, 234.

SIF Sentence Vector Model Based on Word Selection by Variance Weight Factor

SUN Yi, QIU Hangping, KANG Ruizhi

(Institute of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210000, China)

[Abstract] To address the poor performance of the Smooth Inverse Frequency (SIF) model in text classification and sentiment analysis, based on the SIF model, the Variance Weight (VW) of the task contribution is calculated according to the distribution of words in different classification task, and a VW factor selection sentence vector model CwVW-SIF is established. Tested on the standard text classification datasets and sentiment analysis datasets, the results show that CwVW-SIF has higher classification accuracy than SIF model.

[Key words] Smooth Inverse Frequency (SIF); sentence vector; Variance Weight (VW); text classification; emotion analysis

DOI: 10.19678/j.issn.1000-3428.0052381

0 概述

使用不同方法生成词向量是自然语言处理(Natural Language Processing, NLP)和信息检索(Information Retrieval, IR)领域的基本任务之一。目前,词向量生成模型主要有 Word2Vec^[1]、GloVe^[2]、FastText^[3]、PSL^[4]和 ELMo^[5]。其中,Word2Vec 和 GloVe 是基于分布假设的无监督方法,FastText 在 Word2Vec 的基础上添加了基于字符的 N -gram 模型,可以计算表外单词的向量,PSL 模型利用解释数据集 PPDB^[6]有监督地对词向量进行调整,在文本相似性任务上具有较好的性能,ELMo(Embeddings from Language Models)模型利用深度上下文单词表征方法,

学习不同上下文的词汇多义性。

近年来,国内外学者研究了句向量模型在文本相似度比较、文本分类和文本情感分析等下游任务中的应用。文献[7]在 Word2Vec 基础上,提出分布式记忆句向量(Distributed Memory of Paragraph Vector, PV-DM)和分布式词袋句向量(Distributed Bag of Words of Paragraph Vector, PV-DBOW)2 种模型。文献[8]提出神经词袋(Neural Bag-of-Words, NBOW)和深度平均网络(Deep Averaging Network, DAN)2 种句向量模型,实验结果验证了深层无序组合方法的有效性。文献[9]提出 Skip-thoughts 模型,通过训练 2 个循环神经网络(Recurrent Neural Network, RNN)组成的编码-解码模型得到句向量,并

基金项目: 江苏省自然科学基金(BK20150721, BK20161469);江苏省重点研发计划(BE2015728, BE2016904, BE2017616)。

作者简介: 孙 毅(1993—),男,硕士研究生,主研方向为自然语言处理、网络通信;裘杭萍,教授、博士;康睿智,博士研究生。

收稿日期: 2018-08-13 **修回日期:** 2018-09-13 **E-mail:** sunyi_lgdx@sina.com

通过词汇扩展方法来编码训练集外的单词。文献[10]提出 RNNs 模型, 利用长短时记忆 (Long Short-Term Memory, LSTM) 来捕捉长距离依存关系。文献[11]提出 PP (Paragram-Phrase embeddings) 模型, 通过将句子中词的词向量进行算术平均得到句向量, 并利用投影方法来对模型进行优化, 同时运用 PSL 词向量来改善模型在各项任务中的性能。文献[12]采用 TF-IDF 加权的方法形成句向量, 并在文本相似度任务上取得较好的效果。文献[13]提出平滑反频率 (Smooth Inverse Frequency, SIF) 模型, 该模型与 PP 模型相似, 但是选择了加权平均的方法, 并通过移除句子的第一主成分上矢量的方法进行优化, 该方法在各项任务上 (除情感分类任务) 均优于其他方法的性能。文献[14]提出 p-mean 模型, 通过集成学习的方法来提升句向量的性能。

SIF 模型统计了通用数据集上词的频率, 但未考虑与任务无关词的筛选或权重的修正, 在情感分析方面相对 RNN 和 LSTM 方法性能较差。为此, 本文利用方差选词的方法对 SIF 模型进行优化, 去除对分类任务贡献值较低单词, 以提高 SIF 模型在文本分类和情感分析方面的性能。

1 平滑反频率句向量模型

随机游走 (Random Walk, RW) 是网络图的经典算法之一, 从给定图的初始位置出发, 随机地选择并移动到邻居节点上, 将当前节点作为出发点, 迭代上述过程, 其特点是无后效性, 即基于过去的表现, 无法预测将来事件的发生步骤和方向。

平滑反频率模型将语句的产生视为一个动态的随机游走过程, 在第 t 步产生第 t 个单词, 每一步都由一个话题向量 $\mathbf{c}_t \in \mathbb{R}^d$ 决定。对于给定的句子 s , 其句向量是对决定该句子的话题向量 \mathbf{c}_t 的最大后验概率估计。同时, 由于在一句话中话题向量 \mathbf{c}_t 的改变很小即一个句子中的话题相对固定, 因此将所有都近似为 \mathbf{c}_s 。在平滑反频率模型中平滑基于以下2种假设:

- 1) 部分单词并不是根据上下文出现的。
- 2) 一些高频词汇 (如 “the” “and”) 的出现与句子的话题无关。

单词 w 出现在以 \mathbf{c}_s 为话题的句子中的概率为:

$$\Pr[\mathbf{w}_s | \mathbf{c}_s] = \alpha p(w) + (1 - \alpha) \frac{\exp(\langle \tilde{\mathbf{c}}_s, \mathbf{v}_w \rangle)}{Z_{\tilde{\mathbf{c}}_s}} \quad (1)$$

其中, $\tilde{\mathbf{c}}_s = \beta \mathbf{c}_0 + (1 - \beta) \mathbf{c}_s$, \mathbf{c}_0 与 \mathbf{c}_s 正交。第1项 $\alpha p(w)$ 对应假设1, $p(w)$ 表示单词在整个语料集中出现的频率, α 为常量, 允许单词的概率极小, 但仍以 $\alpha p(w)$ 的概率出现。第2项对应假设2, 假设对所有的句子都有一个共同的话题向量 $\mathbf{c}_0 \in \mathbb{R}^d$, 当单词 w 是高频词即与共同话题 \mathbf{c}_0 相关时, 能以一定的概率出现, β 为常量,

$Z_{\tilde{\mathbf{c}}_s} = \sum_{w \in V} \exp(\langle \tilde{\mathbf{c}}_s, \mathbf{v}_w \rangle)$ 将第2项进行归一化。

基于以上的假设, 以 \mathbf{c}_s 为话题的句子 s 的生成概率为:

$$p[s | \mathbf{c}_s] = \prod_{w \in s} p(w | \mathbf{c}_s) = \prod_{w \in s} \left[\alpha p(w) + (1 - \alpha) \frac{\exp(\langle \mathbf{v}_w, \tilde{\mathbf{c}}_s \rangle)}{Z} \right] \quad (2)$$

$$\ln f_w(\tilde{\mathbf{c}}_s) = \ln \left[\alpha p(w) + (1 - \alpha) \frac{\exp(\langle \mathbf{v}_w, \tilde{\mathbf{c}}_s \rangle)}{Z} \right],$$

对其进行微分, 有:

$$\nabla f_w(\tilde{\mathbf{c}}_s) = \frac{1}{\alpha p(w) + (1 - \alpha) \exp(\langle \mathbf{v}_w, \tilde{\mathbf{c}}_s \rangle) / Z} \times \frac{1 - \alpha}{Z} \exp(\langle \mathbf{v}_w, \tilde{\mathbf{c}}_s \rangle) \mathbf{v}_w \quad (3)$$

根据泰勒展开, 有:

$$f_w(\tilde{\mathbf{c}}_s) \approx f_w(0) + \nabla f_w(0)^T \tilde{\mathbf{c}}_s = C + \frac{(1 - \alpha) / (\alpha Z)}{p(w) + (1 - \alpha) / (\alpha Z)} \langle \mathbf{v}_w, \tilde{\mathbf{c}}_s \rangle \quad (4)$$

因此, 对 $\tilde{\mathbf{c}}_s$ 的最大后验估计为:

$$p[s | \mathbf{c}_s] = \sum_{w \in s} f_w(\tilde{\mathbf{c}}_s) \propto \sum_{w \in s} \frac{a}{p(w) + a} \mathbf{v}_w \quad (5)$$

其中, $a = \frac{1 - \alpha}{\alpha Z}$ 。

定义单词 w 在句子 s 中对应的权重为:

$$Weight(w) = \frac{a}{a + p(w)} \quad (6)$$

句子 s 的句向量 \mathbf{v}_s 为:

$$\mathbf{v}_s = \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} \mathbf{v}_w = \frac{1}{|s|} \sum_{w \in s} Weight(w) \mathbf{v}_w \quad (7)$$

句向量 \mathbf{v}_s 是以 $a / (a + p(w))$ 为权重的词向量的加权平均, 根据单词频率 $p(w)$ 的分布规律, a 在 $[10^{-3}, 10^{-4}]$ 范围内对权重的区分度最大, 即在这个范围之外, 不同单词的权重 $a / (a + p(w))$ 基本相等。

2 方差选词方法

在平滑反频率模型中, 随机游走能够较好地反映在通用语料的统计规律下句子的生成规律, 但在具体分类任务中, 没有考虑每一类话题对句子生成的影响。

假设在分类任务中的总体语料数据集为 D , 每一类语料为 $D_i, i \in [2, n]$, 则模型共同的话题向量 \mathbf{c}_0 为总体语料的共同话题, 而针对每一类的语料, 应该有属于该类语料的专有共同话题, 定义为 \mathbf{c}_i , 其中 i 与分类语料对应。

为保持平滑反频率模型的通用性, 同时提高其在分类任务中的准确性, 本文在该模型中添加了方差选词组件, 通过在计算句向量时去除存在共同话题的词, 提升句向量在不同类别中的区分度。

2.1 方差因子计算

设单词 w 在第 i 类语料中出现的概率为:

$$P(w|D_i) = \frac{|\{s|w \in s, s \in D_i\}|}{|D_i|} \quad (8)$$

其中, $|\{s|w \in s, s \in D_i\}|$ 为含有单词 w 的句子的个数, $|D_i|$ 为该语料中所有句子的个数。

无论句子 s 中单词 w 出现过多少次,都记为 1。定义单词 w 在不同类别语料中的方差因子为 $Var(w)$, 则均方差为:

$$S^2(w) = \frac{1}{n-1} \sum_{i=1}^n (X_i(w) - \bar{X})^2 \quad (9)$$

其中, $X_i(w) = P(w|D_i)$ 。

为方便不同单词方差因子的比较, 本文将均方差进行归一化处理, 得到如式(8)所示的方差因子。

$$Var(w) = S^2(w) = \frac{1}{n-1} \sum_{i=1}^n \frac{(X_i(w) - \bar{X})^2}{\bar{X}^2} \quad (10)$$

方差因子越小, 表示该单词的意思在不同类别语料中出现的概率越接近, 即可能属于总体语料的共同话题; 方差因子越大, 表示该单词的意思在不同类别语料中出现的概率相差越大, 即可能属于不同类别语料的专有共同话题。因此, 方差因子越大对分类问题的贡献率越大, 因子越小对分类问题贡献率越小。方差因子较小的单词, 在句向量生成过程中会成为影响分类效果的“噪声”。

为体现方差因子在多分类问题(即 $n \geq 3$) 对类与类之间的区别的贡献度, 本文选取两两类别归一化方差的最大值作为多分类情况下的方差因子, 可表示为:

$$Var(w) = \max_{i,j \in n, i \neq j} S_{i,j}^2(w) \quad (11)$$

2.2 基于方差权重的 SIF 句向量模型

上述方差因子的计算仅考虑单词 w 在具体任务的总体语料数据集 D 中的重要性, 为综合考虑单词 w 在通用背景下的重要性, 本文通过方差权重 (Variance Weight, VW) 因子表征单词对分类任务的重要程度, 即:

$$VW(w) = Var(w) Weight(w) \quad (12)$$

在句向量的生成过程中, 将方差权重因子 $VW(w)$ 去除, 再进行句向量计算, 因此本文提出基于方差权重选词的 SIF 句向量模型 CwVW-SIF, 具体算法描述如下:

算法 1 基于方差权重选词的平滑反频率句向量生成

输入 词向量集 $\{v_w : w \in V\}$, 句子集合 S , 参数 a , 通用语料库统计的单词频率 $\{p(w) : w \in V\}$, 分类任务训练集 $\{D_i : D_i \in D\}$

输出 单词出现频率集 $\{f(w|D_i) : w \in V, D_i \in D\}$, 单词

方差权重因子 $\{VW(w) : w \in V\}$, 句向量集 $\{v_s : s \in S\}$, 单词剪裁数量 k

```

1. for all w in V do
2. f(w|Di) ← get_frequency(w)
3. end for
4. for all w in V do
5. Var(w) ← calculate_var(w)
6. VW(w) ← Var(w) *  $\frac{a}{a + p(w)}$ 
7. end for
8. V_sorted sort V by VW(w)
9. k ← cut_number(V_sorted, D)
10. for all s in S do
11. s ← remove V_sorted[0, k] ins
12. vs ←  $\frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w$ 
13. end for

```

算法 1 的第 1 行和第 2 行表示每个单词在不同类句子或文章中出现的概率, 第 5 行用来计算方差因子, 第 6 行用于方差权重因子的计算, 第 8 行将单词表 V 按照方差权重因子 $VW(w)$ 由小到大排序为单词表 V_sorted , 第 9 行用来计算最佳单词剪裁数, 如算法 2 所示, $cut_number(V_sorted, D)$ 函数通过按照循序依次从训练集中去除单词表 V_sorted 中前 k 个单词, 然后将训练后的句向量用于分类器, 寻找最佳的准确率对应的剪裁数值, 并将其返回, 第 11 行将单词表中的单词去掉, 第 12 行通过加权平均, 将词向量加权平均为句向量。

算法 2 最佳单词剪裁数量计算 $cut_number(V_sorted, D)$

输入 词向量集 $\{v_w : w \in V\}$, 句子集合 S , 参数 a , 通用语料库统计的单词频率 $\{p(w) : w \in V\}$, 分类任务训练集 $\{D_i : D_i \in D\}$, 按照方差权重因子排序的词汇表 V_sorted

输出 准确率集 $ACC = \{acc(i) | i \in [0, len(V_sorted)]\}$, 单词剪裁数 k

```

1. for i in range [0, len(V_sorted)] do
2. for all s in D do
3. s ← remove V_sorted[0, i] ins
4. vs ←  $\frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w$ 
5. end for
6. acc(i) ← SVM( $\{v_s : s \in S\}, D$ )
7. end for
8. return max_number(ACC)

```

3 CwVW-SIF 模型

3.1 句向量分类模型

在句向量模型上加入有监督的分类器, 构成基于句向量的分类模型, 如图 1 所示。首先将分类任

务中带有标记的训练语料输入到 CwVW-SIF 模型, 得到带有标记的句向量, 每个句向量是 m 维的数值向量, 然后将带有标记的句向量输入到分类器中, 经过训练的分类器便可用于测试语料的分类。

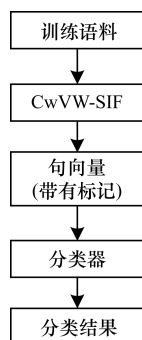


图1 句向量分类模型

分类器即分类算法, 如支持向量机、各类决策树 (如随机森林、极端随机树)、BP (如前馈神经网络) 算法等, 分类器的输入为数值特征向量, 输出为输入数据的分类标记。分类器的选择在整个分类模型中并不起决定性作用, 但通过合理选择分类器, 并对分类器的参数进行调整, 可在一定程度上改善分类效果。

3.2 文本关键词抽取

CwVW-SIF 模型生成句向量的过程是对词向量进行加权求和的过程, 因此, 可利用该过程确定句子中的关键词。某一个单词 w 的权重值可表示为方差权重因子, 即:

$$FVW(w) = F(w) Var(w) Weight(w) \quad (13)$$

其中, $F(w)$ 表示单词在句子中出现的次数, 即某一单词在同一句子中出现的次数越多, 对该句子的贡献越大, $Var(w)$ 表示单词 w 的归一化方差因子, 其取值范围为 $[0, 1]$, 越接近 1 对分类的贡献越大, 对句子的中心主题描述的贡献就越大, $Weight(w)$ 表示单词 w 的权重因子, 与单词的统计频率呈反比, 即单词在全体文本中出现的频率越大, 对句子含义的表的贡献值越小。通过方差权重因子, 计算句子中单词所对应的 FVW 值, 并根据排序, 筛选出句子的关键词。

4 实验结果与分析

4.1 数据集

本文使用公开分类任务数据集 20 Newsgroups^[15-16]和取自 IMDB 的情感分析任务数据集 Large Movie Review Dataset^[17-18]进行实验。

数据集 20 Newsgroups 主要用于文本分类、文本挖掘和信息检索研究, 共收录 20 个不同主题的新闻约 20 000 篇, 训练集和测试集分别占 60% 和 40%, 分类情况如表 1 所示。其中部分新闻类型极为相似 (如 comp. sys. ibm. pc. hardware 和 comp. sys. mac. hardware), 也有一些类别之间完全不同 (如 misc. forsale 和 soc. religion. christian)。

表1 数据集 20 Newsgroups 类型

序号	类型
1	comp. graphics
	comp. os. ms-windows. misc
	comp. sys. ibm. pc. hardware
	comp. sys. mac. hardware
2	comp. windows. x
	rec. autos
	rec. motorcycles
	rec. sport. baseball
3	rec. sport. hockey
	sci. crypt
	sci. electronics
	sci. med
4	sci. space
	misc. forsale
5	talk. politics. misc
	talk. politics. guns
	talk. politics. mideast
6	talk. religion. misc
	alt. atheism
	soc. religion. christian

数据集 Large Movie Review Dataset 是通用的情感二分类数据集, 共有 50 000 条源自 IMDB 的评论, 训练集和测试集各有 25 000 条样本, 正负样本各 12 500 条。

本文采用的词向量集为 glove. 6B. 50d^[19], 该数据集是在维基百科语料库上根据 GloVe 模型训练得到, 共有 40 万个单词, 每个单词表示为 50 维的向量, 选择该向量集有如下原因: 1) 该向量集由斯坦福大学训练并公开, 较为成熟, 具有通用性和可比性; 2) 该训练集将单词表示为 50 维度的向量, 模型的训练速度相对较快。

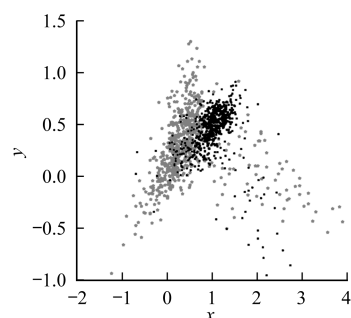
单词频率数据集 enwiki_vocab_min200 是由维基百科语料统计而来^[20], 共含有 34.8 万个单词。

4.2 模型效果的 PCA 降维

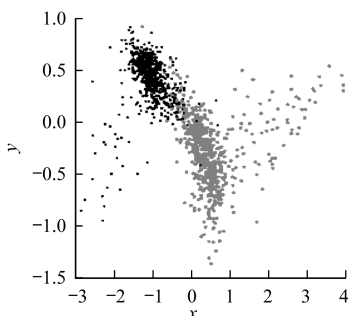
在通过分类器对改进的 CwVW-SIF 模型进行性能度量之前, 本文运用改进前和改进后的模型, 分别通过主成分分析方法, 将句向量投影到低维空间进行可视化效果展示并进行比较。

本文选取 2 对数据集 (大差别、小差别) 进行实验, 经过 SIF 模型和 CwVW-SIF 模型将每一条新闻文本转化为 50 维的句向量, 再将句向量通过 PCA 降维到 2 维和 3 维进行观察。

大差别数据集选取 20 Newsgroups 中类别之间有较大差别的 comp. graphics 和 soc. religion. christian 这 2 类数据进行展示。其中, comp. graphics 类 584 条, soc. religion. christian 类 599 条。其二维效果如图 2 所示, 三维效果如图 3 所示, 灰色的点为 comp. graphics 类, 黑色的点为 soc. religion. christian 类。可以看出, 经过优化后的模型训练出的数据, 同一类更加紧凑, 不同类之间区分更加明显。

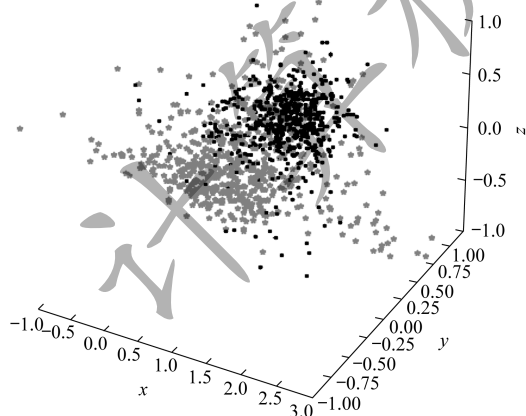


(a)未经优化的模型

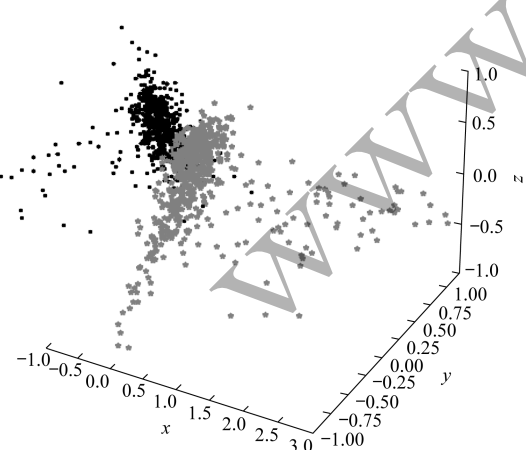


(b)经过优化的模型

图2 大差别数据集二维效果展示



(a)未经优化的模型

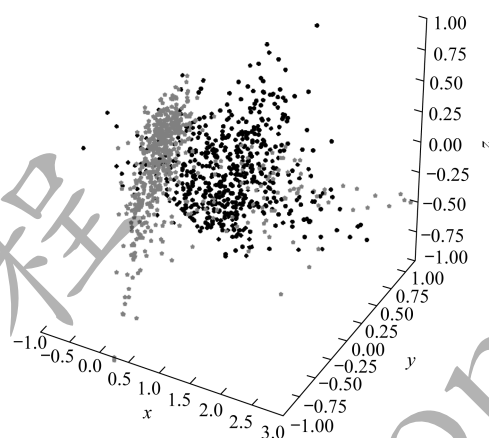


(b)经过优化的模型

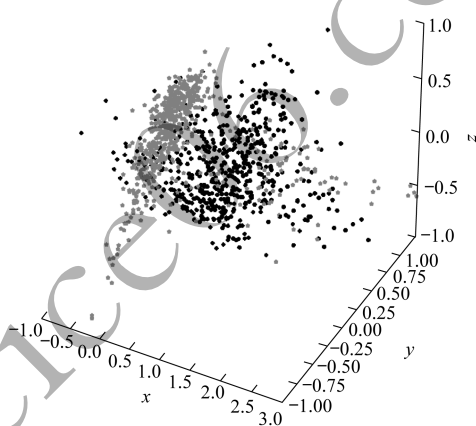
图3 大差别数据集三维效果展示

小差别数据集选取了 20 Newsgroups 中类别之间极为相似的 comp. sys. ibm. pc. hardware 和 comp.

sys. mac. hardware 两类数据进行展示。其中, comp. sys. ibm. pc. hardware 类 590 条, comp. sys. mac. hardware 类 578 条。由于 2 个数据集差别较小, 因此三维图展示效果如图 4 所示, 其中灰色的点为 comp. sys. ibm. pc. hardware 类, 黑色的点为 comp. sys. mac. hardware 类。



(a)未经优化的模型视角



(b)经过优化的模型视角

图4 小差别数据集三维效果

4.3 模型分类效果

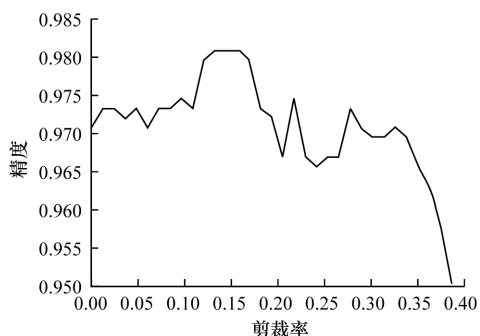
本文重点在于对比模型改进前后的效果, 因此在分类器上选择支持向量机, 且只简单对其参数进行调节(情感分析任务同样如此)。

支持向量机输入由训练集经 SIF 模型和 CwVW-SIF 模型产生的句向量, 输出为分类结果。其采用高斯核函数, 核函数系数为 5, SVC 的惩罚值为 3, 停止训练误差为 10^{-3} , 无最大迭代次数限制, 决策函数为 OVR。根据调试, 权重计算参数 $a = 2.7 \times 10^{-3}$ 时分类效果最佳。

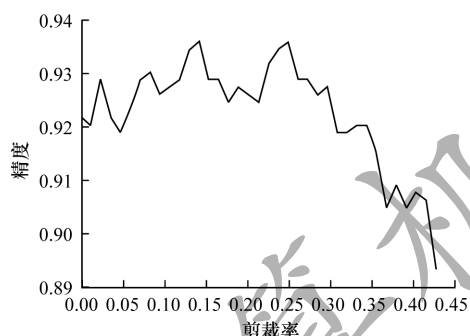
CwVW-SIF 模型在分类任务中最重要的是找到最佳的单词剪裁率, 即算法 2 所示的最佳单词剪裁数量。以不同的比例除去按照方差权重因子排序后的单词表的部分单词, 通过训练集对模型进行训练, 然后根据验证集找到最佳的剪裁率, 并在验证集上检测模型效果。

4.3.1 文本分类任务

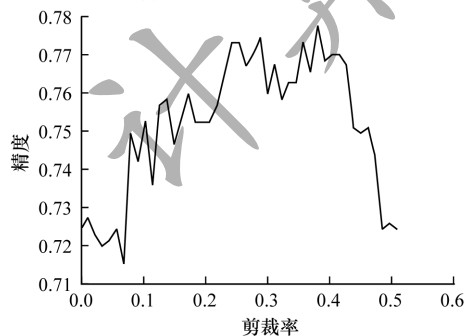
本文选取 20 Newsgroups 中难度不同的 4 种话题类型任务对本文模型进行实验, 其精度结果如图 5 所示。



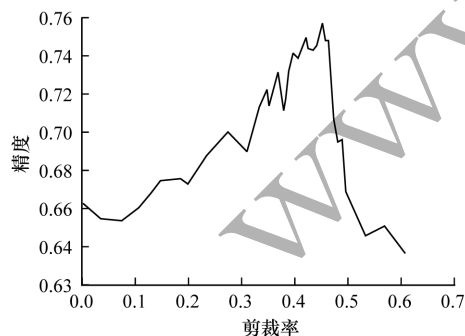
(a) comp.graphics/soc.religion.christian 分类任务



(b) alt.atheism/sci.med 分类任务



(c) talk.politics.misc/talk.politics.guns 分类任务



(d) comp.sys.ibm.pc.hardware/comp.sys.mac.hardware 分类任务

图 5 不同分类任务下模型精度对比

表 2 给出不同分类任务精度和最佳剪裁率对比结果。可以看出, 4 种任务的最佳剪裁率分别为 14.5%、24.8%、38.2%、45.4%, 对于每个分类任务, 在剪裁率从 0% 递增到 100% 的过程中, 算法的精度先提升, 达到极

值后, 再下降, 且任务难度越大, 最佳剪裁率越大。这是由于分类任务难度大, 主题无关或与分类无关的词汇多, 对于句向量的影响就越大。对于类别 comp. sys. ibm. pc. hardware 与类别 comp. sys. mac. hardware, 两者均为计算机硬件领域, 前者是 IBM 公司, 后者是苹果公司, 只有少数的关键词才会对分类起到决定性的作用。因此通过方差因子去除无关词汇, 再进行句向量生成, 能够提高分类任务的性能。

表 2 不同分类任务精度和最佳剪裁率对比

分类任务	精度		精度提升	最佳剪裁率
	SIF 模型	CwVW-SIF 模型		
comp. graphics/ soc. religion. christian	0.970	0.981	0.011	0.145
alt. atheism/sci. med	0.921	0.935	0.014	0.248
talk. politics. misc/ talk. politics. guns	0.724	0.777	0.053	0.382
comp. sys. ibm. pc. hardware/ comp. sys. mac. hardware	0.663	0.757	0.094	0.454

4.3.2 情感分析任务

本文情感分析任务选取数据集 Large Movie Review Dataset 中不同数据规模的数据, 数据规模分别为 1 000 条、2 500 条、5 000 条、10 000 条和 20 000 条, 利用 CwVW-SIF 模型找到单词表的最佳剪裁率, 得到模型的最佳性能, 分类器选择支持向量机。

支持向量机输入由训练集经 SIF 模型和 CwVW-SIF 模型产生的句向量, 输出为分类结果。支持向量机采用高斯核函数, 核函数系数为 3, SVC 的惩罚值为 20, 停止训练误差为 10^{-3} , 无最大迭代次数限制, 决策函数为 OVR。根据调试, 权重计算参数 $a = 2.7 \times 10^{-3}$ 时分类效果最佳。

2 种模型在 5 种训练集规模下的精度如表 3 所示。不同剪裁率对 CwVW-SIF 模型精度的影响如图 6 所示。可以看出, 5 种训练规模下 CwVW-SIF 模型对于 SIF 模型都有提高, 且随着训练规模的增大, 性能提升幅度也随之增大。

表 3 不同规模训练集的性能对比

训练集规模	精度		精度提升	最佳剪裁率
	SIF + SVM 模型	CwVW-SIF + SVM 模型		
1 000	0.746	0.760	0.014	0.230
2 500	0.750	0.779	0.029	0.260
5 000	0.757	0.791	0.034	0.360
10 000	0.768	0.815	0.047	0.440
20 000	0.772	0.813	0.036	0.400

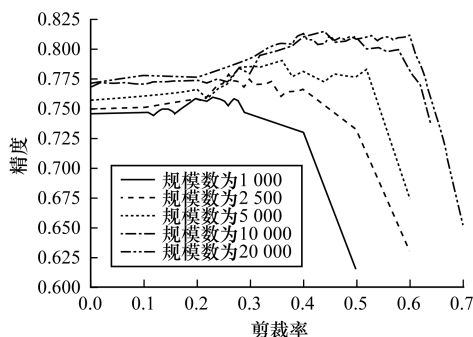


图 6 5 种训练集规模下精度对比结果

4.4 关键词提取

本文通过对数据集 20 Newsgroups 的文章进行关键词抽取,检验方差权重因子的效果,选取 comp. sys. ibm. pc. hardware (类别 0)、rec. sport. baseball (类别 1) 2 种类别的文章,然后对比文章题目和根据本文算法抽取的 Top5 关键词。同时,由于关键词提取不是本文的重点,且数据集 20 Newsgroups 中文章的题目并非其关键词,因此只举例说明,不做命中率统计,结果如表 4 所示。其中,加粗表示题目中含有的关键词和方差权重因子,且不区分大小写。可以看出,本文算法对关键词的抽取效果较好。

表 4 关键词提取效果展示

文章题目	类别	关键词/FVW 权重
NetBIOS and BIOS	0	netbios/4.00, emilio/1.95, bios/1.95, interrupt/0.99, sci/0.96
IDE vs SCSI	0	scsi/23.86, pc/10.57, asynchronous/7.94, ide/7.90, mac/7.86
Re:AMD i486 clones: Now legal in US?!?!?!?	0	amd/2.99, clones/2.94, sward/2.00, prohibiting/1.97, chip/1.86
Ram boards on a 486??	0	ram/5.15, simms/4.95, oracle/2.88, isa/2.69, boards/2.68
ide & scsi controller	0	ram/3.43, hd/2.73, scsi/1.99, ide/1.96, cache/1.89
WHAT'S WITH ALL THESE SCORES?	1	hernandez/3.91, scores/2.57, mailing/2.45, posts/1.74, standings/1.71
WFAN	1	wip/8.97, lupica/4.00, berman/3.94, sports/3.94, fan/2.81
Hal McRae	1	mcrac/3.96, rbd/2.00, mcgraw/1.95, royals/1.90, davis/1.32
Let's Talk Phillies	1	phillies/4.73, phils/3.98, cubs/2.82, homers/1.99, myers/1.91
Re:Royals	1	royals/1.90, depressis/1.00, spork/1.00, izzo/1.00, mc Reynolds/1.00

5 结束语

本文根据单词在分类任务中的分布情况,建立基于方差权重选词改进的平滑反频率句向量模型 CwVW-SIF。在文本分类和情感分析 2 种任务上进行实验,结果表明,该模型具有较高的分类精度。由于在单词剪裁率增长的过程中,精度曲线并不完全平滑,因此下一步将优化单词的权重因子来解决该问题。

参考文献

- [1] MIKOLOV T, CORRADO G, CHEN Kai, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-07-10]. <https://arxiv.org/pdf/1301.3781.pdf>.
- [2] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2014: 1532-1543.
- [3] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification [C]//Proceedings of Conference of European Chapter of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017: 427-431.
- [4] WIETING J, BANSAL M, GIMPEL K, et al. From paraphrase database to compositional paraphrase model and back [J]. Transactions of the Association for Computational Linguistics, 2015, 3: 345-358.
- [5] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL]. [2018-07-10]. <https://arxiv.org/pdf/1802.05365.pdf>.
- [6] GANITKEVITCH J, VANDURME B, CALLISON-BURCH C. PPDB: the paraphrase database[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2013: 758-764.
- [7] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning. Cambridge, USA: MIT Press, 2014: 1188-1196.
- [8] IYYER M, MANJUNATHA V, BOYD-GRABER J, et al. Deep unordered composition rivals syntactic methods for text classification [C]//Proceedings of International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 1681-1691.
- [9] KIROS R, ZHU Yukun, SALAKHUTDINOV R, et al. Skip-thought vectors [C]//Proceedings of International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2015: 3294-3302.
- [10] TAI Kaisheng, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks [C]//Proceedings of International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015: 1556-1566.

(下转第 234 页)

(上接第 210 页)

- [11] WIETING J, BANSAL M, GIMPEL K, et al. Towards universal paraphrastic sentence embeddings [EB/OL]. [2018-07-10]. <https://arxiv.org/pdf/1511.08198.pdf>.
- [12] 段旭磊, 张仰森, 孙伟卓. 微博文本的句向量表示及相似度计算方法研究 [J]. 计算机工程, 2017, 43 (5): 143-148.
- [13] ARORA S, LIANG Y, MA Tengyu. A simple but tough to beat baseline for sentence embeddings [EB/OL]. [2018-07-10]. <https://openreview.net/pdf?id=SyK00v5xx>.
- [14] RÜCKLÉ A, EGER S, PEYRARD M, et al. Concatenated p-mean word embeddings as universal cross-lingual sentence representations [EB/OL]. [2018-07-10]. <https://arxiv.org/pdf/1803.01400.pdf>.
- [15] LANG K. NewsWeeder: learning to filter netnews [C]// Proceedings of International Conference on International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1995: 331-339.
- [16] 20 newsgroups [EB/OL]. [2018-07-10]. <http://www.qwone.com/~jason/20Newsgroups>.
- [17] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis [C]// Proceedings of Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2011: 142-150.
- [18] Large movie review dataset [EB/OL]. [2018-07-10]. <http://ai.stanford.edu/~amaas/data/sentiment>.
- [19] GloVe: global vectors for word representation [EB/OL]. [2018-07-10]. <https://nlp.stanford.edu/projects/glove>.
- [20] Wikimedia. English Wikipedia dump [EB/OL]. [2018-07-10]. <http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2>.

编辑 赵 辉