基于 word2vec 的文本情感分析方法研究

◆彭晓彬

（广东工业大学计算机学院　广东　510006）

摘要：word2vec 是一个将单词转换成向量形式的工具，可以把对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的相似度，来表示文本语义上的相似度。本文通过对 word2vec 工作原理的介绍并将它应用在文本情感分析领域，最后结合一些分类器取得较好的实验结果，相较于传统的文本情感分析方法（例如基于情绪词表的方法）在预测情感分类的准确率上有所提高。

关键词：word2vec；词向量；情感分析；NLP

0　引言

情感分析的主要目的就是识别用户对事物或人的看法、态度，参与主题主要包括：观点持有者、评价的对象、评价的观点，评价的文本（一般是一个句子或者是整篇文档）[1]。

本文主要是将 word2vec[2]应用于情感分析领域，结合一些分类器来做实验，并与传统的 0-1 语言模型对比。

1　相关介绍

1.1　词袋模型

“词袋模型”[3]是在自然语言处理领域的一种语言模型。该模型将文本转化为基于词典的 0-1 向量，并认为词语之间没有语义上的联系，只考虑词语的共现关系，忽略了上下文的信息和词语之间的语义关联。

1.2　word2vec 的原理

Word2Vec 算法可以在捕捉语境信息的同时压缩数据规模。其中，谷歌提供的 word2vec 工具包含了 CBOW 和 Skip_gram 两种语言模型[4]，这两个模型均包含输入层、投影层和输出层。其中，CBOW 模型（如图 1 所示）通过上下文来预测当前词，与其相反，Skip_gram 模型（如图 2 所示）则通过当前词来预测其上下文窗口内的单词。Word2vec 提供了两种优化方法来提高词向量的训练效率，分别是 HierachySoftmax 和 Negative Sampling。这些方法都利用人工神经网络作为它们的分类算法[5]。每个单词都是一个随机 N 维向量，经过训练之后获得了每个单词的最优向量。

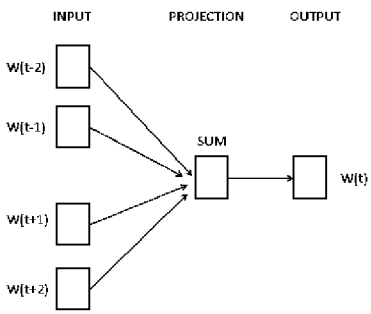

图 1　CBOW 的语言模型结构

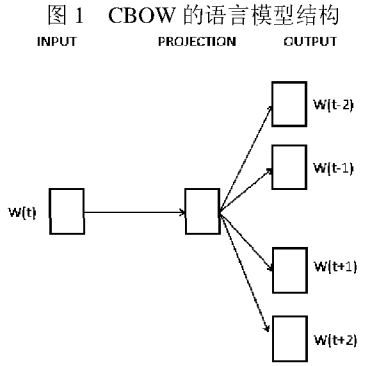

图 2　skip-gram 的语言模型结构

学习某一个词的词向量的框架如图 3 所示，该框架是在给定上下文的词语的条件下来预测下一个词。该框架中，词语被投射在一个向量空间中，每一个词语对应矩阵 W 里面唯一的一个列向量，以词语在词汇表里面的位置为索引编号[6]，然后上下文的单词向量的级联或加和作为特征向量来预测句子中的下一个词语。在图 2 的框架中，用前面三个单词“the”、“cat”、“sat”来预测第四个单词“on”，每一个单词被影射到向量矩阵 W 中。

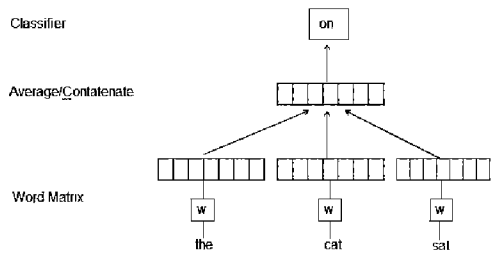

图 3　学习词向量的框架

当给定一个需要训练的词的序列 w_1，w_2，w_3，…，w_T，词向量模型的目标是最大化概率取 log 的平均值

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t \mid w_{t-k},\ldots,w_{t+k}) \qquad ①$$

这个预测任务主要通过利用类是多分类的做法，例如 softmax，在上式中，后验概率

$$p(w_t \mid w_{t-k},\ldots,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum e^{y_i}} \qquad ②$$

每个 y_i 都是没经过归一处理概率的 log 值，可以做以下计算。

$$y = b + Uh(w_{t-k},\ldots,w_{t+k};W) \qquad ③$$

U 和 b 是 softmax 的参数，h 可以通过词向量的加权平均值进行构建。

通过 word2vec 算法模型，现在这些词向量已经捕捉到上下文的信息。我们可以利用基本代数公式来发现单词之间的关系。比如存在数学关系 C（king）−C（queen）≈C（man）−C（woman），与 C（king）−C（man）+C（woman）最接近的向量就是 C（queen）。[7, 8]这些词向量可以代替词袋用来预测未知数据的情感状况。该模型的优点在于不仅考虑了语境信息，还压缩了数据规模。由于文本的长度各异，我们可能需要利用所有词向量的平均值作为分类算法的输入值，从而对整个文本进行分类处理。

2　实验

我们利用 word2vec 的 python 语言版本，可以直接训练自己语料库的词向量或者直接导入已经训练好的词向量。实际上，当用于训练的原始语料规模越大，所得词向量质量就越高，利用谷歌预训练好的词向量数据来构建模型是非常有用的，该词向量是基于谷歌新闻数据（大约一千亿个单词）训练所得。这个文件解压后的大小是 3.5 GB（数据来源 https://code.google.com/p/word2vec/）。

2.1 单句子短文本评论集的情感分析

本实验中，用于情感分类的数据集是某网上商城的商品评论集。该数据集拥有约 10000 个短文本，每个文本是一个单独的句子并且带有一个人工标注的标签，positive 表示正面评论，negative 表示负面评论，其中正面评论和负面评论的数量基本对等，本问题可视为二分类问题处理。

实验首先随机从原始数据集中抽取样本，划分训练集和测试集，然后利用 Word2Vec 模型得到训练集的词向量。实验中利用机器学习模块 Scikit-Learn 构建一个线性模型。分类器的选择为 SGDClassifier 中的 SVM，采用随机梯度下降法进行分类器训练，最后利用分类器对测试集进行分类。其中分类器的输入值为推文中所有词向量的加权平均值。实验步骤如下：

（1）导入数据并进行简单的预处理；

（2）利用 sklearn 中的 train_test_split 函数将原始数据集分为训练集和测试集，比例为 8：2；

（3）构建 word2vec 模型，输入训练集数据，获得维度为 300 的词向量；

（4）对于每一条评论文本，计算所有词语的词向量每个维度的的算术平均值，得到的新的向量代表一条文本；

（5）对所有的文本构成的向量矩阵进行归一化，转换后每个维度数据均值为 0，方差为 1；

（6）使用逻辑回归的随机梯度下降法作为分类器算法进行模型训练；

（7）计算测试集的预测精度，构建 ROC 曲线来验证分类器的有效性。

ROC 曲线如下图：

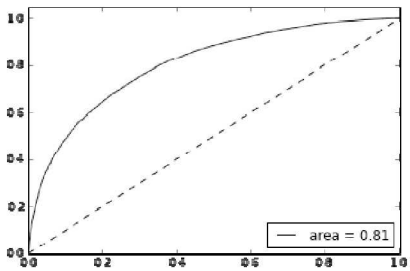

图 4 基于商品评论集构建的分类器的 ROC 曲线

ROC 曲线是评价二类分类器预测精度的一种常用方法，其中横轴为 FTR（假正率），纵轴为 TPR（真正率），绘制曲线的数值由 sklearn 中的 roc_curve 函数生成。当曲线下方的面积越接近 1，表明分类器的分类效果越好。我们利用 Scikit-Learn 构建的简单线性模型的预测精度为 76%，这说明 Word2Vec 模型有效地保留了文本的语义信息，和一个简单的线性分类器的结合较好地对本数据集作出了褒贬分类，取得较好的效果。

2.2 与传统情感分析方法的对比

为了将本文提出的情感分析方法与传统方法进行对比，我们设计了一个基于词袋模型的实验。词袋模型（Bag Of Words，BOW）假设文本中的词语相互独立，词语之间没有语义上的关联，并且认为词语与类别之间是直接关联的。词袋模型在对文本转化为向量表示时，根据特征选择的单词构成的词典将文本映射为一个 0-1 向量。

数据集依然用之前采用的商品评论集，情感分类有褒义和贬义两个类别。特征选择采用 x^2 统计量（CHI）方法，用来计算某个词 t 与类别 c 之间的关系，CHI 值计算公式为：

$$x^2(t,c)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中，A——词 t 与类别 c 共现的文档数；

B——词 t 出现而类别 c 未出现的文档数；

C——词 t 未出现而类别 c 出现的文档数；

D——词 t 与类别 c 均未出现的文档数；

N——文档总数；

依据上式分别计算出训练集中两个类别每个词的 CHI 值，取其中较大者作为词 t 的 CHI 值，即

$$x^2_{max}(t)=max[x^2(t,c_i)]$$

将所有的的词根据 CHI 值进行排序，分别取前 2500、5000、7500、10000 个单词作为特征进行实验，训练分类器分别采用朴素贝叶斯分类器和 SVM 分类器。实验过程采用 10 折交叉验证法，准确率作为评价标准，取十次实验的平均值，不同的特征数量的准确率如下表所示：

表 1 基于词袋模型的分类准确率对比

特征数	特征选择方法	朴素贝叶斯	支持向量机
2500	CHI	0.474	0.581
5000	CHI	0.485	0.593
7500	CHI	0.501	0.604
10000	CHI	0.487	0.611

实验结果表明，在词袋模型中，随着特征数量的提高，分类准确率有所提高，但也造成了维度灾难等问题。另外，由表中数据可知 SVM 更适合词袋模型。其中最高的准确率是在 10000 维度的特征下由 SVM 分类器取得的，相对于前个实验的预测准确率 86%，很明显基于 Word2vec 的实验结果更好，因为词向量与段落向量保存了语义信息。

3 实验结论

在本文跟词袋模型的对比中可以看到 Paragraph Vector 和词向量的平均的联合对效果提升的优越性。通过算法，我们可以获得丰富的词向量和段落向量，这些向量数据可以被应用到各种各样的 NLP 应用中。Word2vec 训练得到的词向量很好地保存了词语之间的语义信息。目前，词向量在中文领域还有很大的发展空间，应用在中文的自然语言处理有望促进中文自然语言处理领域的发展。

参考文献：

[1]Collobert R,Weston J,Bottou L,et al.Natural Lan-guage Processing(almost) from Scratch[J].Journal of Ma-chine Learning Research,2011.

[2]赵妍妍，秦兵，刘挺.文本情感分析[J].软件学报,2010.

[3]Jaakkola T S,Haussler D.Exploiting Generative Models in Discriminative Classifiers[J].Advances in Neur-alInformation Processing Systems,1998.

[4]Le Q V,Mikolov T.Distributed Representations of Sentences and Documents[J].Eprint Arxiv,2014.

[5]Wallach H M.Topic modeling:beyond bag-of-words[C]//International Conference on Machine Learning.2006.

[6]Pang B,Lee L.Seeing stars：exploiting class rel-ationships for sentiment categorization with respect to r-ating scales[J].Proceedings of the Acl,2005.

[7]唐慧丰，谭松波，程学旗.基于监督学习的中文情感分类技术比较研究[J].中文信息学报，2007.

[8]Wallach H M.Topic modeling:beyond bag-of-words[C]//International Conference on Machine Learning,2006.