# 基于多模态数据的社交平台谣言检测

**作者**：王彬，黄旺辉，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：分类、文本嵌入、多模态学习、机器学习、深度学习

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 由于在微博、今日头条等互联网新闻类APP上存在大量的虚假新闻信息，且传统的人工审核内容的方式不适用于如今开放的大数据互联网时代，因此对谣言的自动化检测变得尤为重要。本案例选取了来源于微博的多模态的谣言检测数据集作为数据挖掘对象，其数据来源于微博上的推文，每一条推文均包含文本信息与图片信息，以及说明推文真实性的标签。在实验部分，本案例设计了多组实验，包括对比传统机器学习算法（贝叶斯分类器、支持向量机）与深度神经网络的对比、使用单模态数据与多模态数据的效果对比、文本特征提取方式与分类算法的搭配对比等。在所有实验中，使用多模态数据的深度网络分类器取得最优性能，其精度（Accuracy）达到了82.60%。

**关键词**：谣言检测、多模态学习、机器学习、深度学习

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为基于多模态数据的社交平台谣言检测，需引导学生进行的主要内容有：（1）了解多模态谣言检测问题定义与数据集形式，对数据集进行筛选处理；（2）选取适当的机器学习或深度学习方法，进多模态谣言检测任务，本案例以贝叶斯分类器、SVM分类器、以及基于深度网络的BERT与VGG融合方法作为参考方法。（3）使用Accuracy、Precision、Recall与F1指标作为评价指标，对所用算法进行对比分析，并分析不同文本嵌入方法与多模态数据对结果的影响。

## 2 背景介绍

在当今的移动互联网时代，微博、今日头条及各大新闻类APP成为了信息传播的重要途径，为人们的生活带来了很大的便利。与此同时，虚假新闻与网络谣言也大量地充斥于互联网上，其无休止的传播会为人们的认知甚至社会的稳定都带来很大危害。谣言或虚假新闻检测的目标为识别出捏造的、可以被证实为虚假的新闻内容。

目前多模态谣言检测任务的难点在于：（1）如何有效提取多模态数据的特征？用于分类问题的机器学习方法有很多，但分类性能的上限由特征的有效性决定，尽管深度学习特征提取方法能有效地自动学习特征，但深度学习方法极度依赖于数据质量，且目前的多模态谣言检测数据集规模较小，深度学习特征提取器难以发挥最大性能；（2）如何有效地融合多模态数据的特征？各模态数据分布不同，存在模态鸿沟，难以有效进行跨模态数据间的关联计算，具体来说，对于文本与图像组成的推文，机器难以验证文本对图像的描述是否准确，相关性是否强。

## 3 内容

**3.1 多模态谣言检测问题的定义**

本案例考虑使用文本与图像两种模态的数据来进行谣言检测任务。用*R*表示一组新闻纪录的集合，每条新闻表示为一个元组，其中*tr*为一段文本，*ir*为一张图片，*lr*为这条新闻的真实性标签，取值为0表示真实可信，取值为1表示虚假。

多模态谣言检测任务为学习到一个分类器C，使用新闻的文本与图像作为输入，得到预测的新闻真实性标签，且尽可能使的。

**3.1 数据集及其预处理**

本案例使用的数据集来源于论文[1]的公开发布，下载链接为[[1]](#footnote-0)。该数据集是一个收集自微博社交平台、包含推文的文本、图像以及社交上下文信息，分为谣言集与非谣言集。谣言集的所有推文来自于微博官方辟谣平台发布的从2012年5月至2016年1月间的谣言推文，微博官方辟谣平台鼓励用户举报疑似谣言的推文，然后由审核专家进行人工审核以筛选出真正的谣言。非谣言集来源于可信官方账号（新华社）确认的推文，以保证推文的真实性。训练集与测试集经过精心地划分，以保证两个集合不包含共同的事件，划分比例为8:2。

为了保证训练后续算法的顺利进行，需要对原始数据集进行预处理，包括去掉文本过短的样本与去掉长宽比例奇怪的图片。表1为处理后数据集的数量统计；图1为两条数据样例。

表 1 微博数据集的数量统计

|  |  |  |
| --- | --- | --- |
|  | 真实推文样本数 | 虚假推文样本数 |
| 训练集 | 3568 | 3615 |
| 测试集 | 939 | 900 |

**3.3 分类算法**

（1）贝叶斯文本分类

本案例采用贝叶斯分类器[2]进行仅使用文本信息的谣言检测任务。

首先这里给出条件概率公式：P( A | B ) = P( B | A ) \* P( A ) / P( B )，A | B表示A给定B的概率，也就是说，如果B发生，A发生的可能性有多大。反之亦然。

贝叶斯公式是在条件概率的公式上发展而来的，条件概率的公式为：



贝叶斯公式是条件概率公式的变型，而贝叶斯公式是整个贝叶斯预测模型行的核心：



|  |  |  |
| --- | --- | --- |
| 文本 | 图片 | 标签 |
| 不要再随便买奶制品给孩子喝了❗️幼儿园都发通知了家长们注意：现在得白血病的小孩越来越多，妇幼保健院提示您，请不要给宝宝喝爽歪歪和有添加剂的牛奶饮料，告诉家里有小孩的朋友，旺仔牛奶、可口可乐、爽歪歪、娃哈哈AD钙奶、未来星、QQ星、美汁源果粒奶优菠萝味的。都含有肉毒杆菌。现在紧急召回。 |  | 1（谣言） |
| 【男子遭遇电信诈骗 银行门口上吊自杀】据河南电视台 @都市报道 ，一个月前，周口的熊先生带着1万元来到新乡做小生意，没想到前天接到电信诈骗电话，把1万块钱通过农行汇给了骗子，受骗后向银行和公安求助。今早，因为想不开，他来到农业银行门口自杀。目前警方已经介入此事。 |  | 0（真实） |

图 1 微博数据集数据样例

（2）SVM分类器

本案例采用SVM分类器[2]进行仅使用文本信息的谣言检测任务。

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，  即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。支持向量机的超平面划分如图2所示。

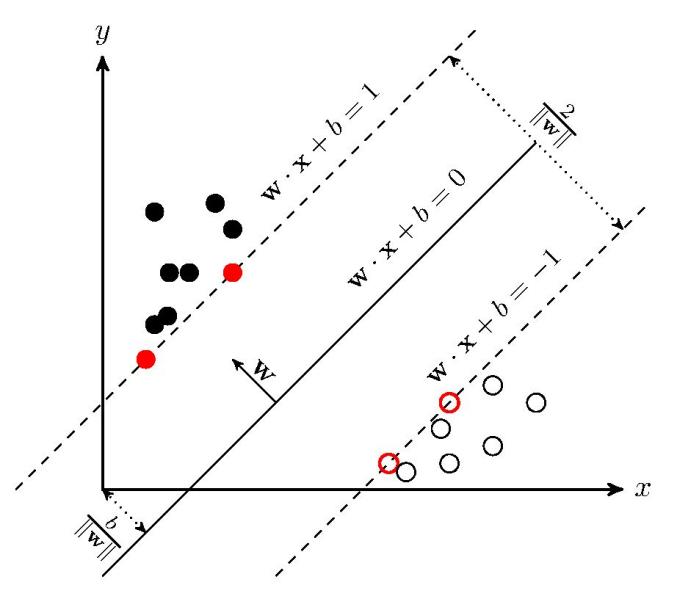


图 2 支持向量机分隔超平面图示

每个支持向量到超平面的距离：



最后得到的最优化问题是：



SVM的分类效果与特征向量的质量有很大关系，在本案例中，使用了TF-IDF（term frequency–inverse document frequency）与Word2Vector[3]两种文本嵌入方法来对文本编码，然后送入SVM分类器学习。TF-IDF考虑词频（Term Frequency）与逆文本频率指（Inverse Document Frequency两种特征来编码文本特征。Word2Vec使用神经网络生成词向量，在本案例试验中，将一段文本的所有词向量取平均得到本案例特征。

（3）多模态的深度神经网络分类

本案例采用深度神经网络分别进行基于单模态数据的谣言检测与基于多模态数据的谣言检测。

由于微博数据集数据量不大，仅为千级，且自然语言及图像特征提取模型参数量大，不易训练，本案例采用了在中文百科数据上预训练的Bert模型提取文本特征，采用在ImageNet上预训练的VGG16模型提取图像特征，提取到的文本特征为768维向量，图像特征为1000维向量。

记一组训练样本为，首先使用预训练模型进行特征提取：



然后分别使用两个双层的全连接网络将图像与文本特征降维至500维，为可训练参数。



接下来将两个向量拼接，送入一个四层的神经网络进行分类，前三层使用ReLU激活函数，最后一层使用Sigmoid激活函数，以保证输出在[0,1]区间内。为可训练参数。



损失函数采用二分类交叉熵损失，定义如下：



其中*l* 为真实标签，为神经网络预测的标签。

使用Adam优化器对网络的可训练参数进行优化，即训练过程可表示为：



对于使用单模态数据的分类器，只需去除相对应的特征降维网络并调整分类器的输入大小即可。

**3.4 实验与分析**

本案例采用的评价指标为准确率（Accuracy)，精度（Precision)，召回率（Recall）与F1分数（F1 Score)，其中准确率作为主要评价指标。

实验细节：（1）BOW、TD-IDF、与Word2Vec嵌入需要首先将文本进行分词，本案例实验采用jieba中文分词工具，并去除停用词。（2）BOW、TD-IDF、与Word2Vec文本嵌入均在整个训练集的所有文本上进行训练，其中BOW嵌入在不限制词典大小时，获得的文本嵌入长度为34517；Word2Vec使用python的Ginsim包提供的方法，对于一段文本，所有单词的词向量进行平均得到文本向量。（3）神经网络使用Pytorch实现，Bert模型采用在中文维基百科上预训练的模型，来源于<https://github.com/ymcui/Chinese-BERT-wwm>, VGG16模型由Pytorch提供的在ImageNet上预训练的模型。

实验结果如表2所示。

表 2 实验结果总览

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **模型** | **数据** | **特征提取** | **Accuracy** | **Precision** | **Recall** | **F1** |
| **bayes** | text | 10000维BOW文本嵌入 | 0.5993 | 0.9218 | 0.1971 | 0.3248 |
| **bayes** | text | 30000维BOW文本嵌入 | 0.6679 | 0.8913 | 0.3652 | 0.5182 |
| **bayes** | text | 34517维BOW文本嵌入（不限制） | 0.6804 | 0.8839 | 0.3987 | 0.5495 |
| **svm** | text | 500维TF-IDF文本嵌入 | 0.6505 | 0.7747 | 0.4020 | 0.5293 |
| **svm** | text | 1000维TF-IDF文本嵌入 | 0.6772 | 0.8224 | 0.4332 | 0.5675 |
| **svm** | text | 5000维TF-IDF文本嵌入 | 0.6870 | 0.8968 | 0.4064 | 0.5594 |
| **svm** | text | 1024维Word2Vec文本嵌入 | 0.5846 | 0.7288 | 0.2394 | 0.3604 |
| **svm** | text | 512维Word2Vec文本嵌入 | 0.5977 | 0.7476 | 0.2673 | 0.3938 |
| **svm** | text | 128维Word2Vec文本嵌入 | 0.6000 | 0.7555 | 0.2684 | 0.3961 |
| **NN** | text | 768维Bert嵌入 | 0.8042 | **0.8253** | 0.7911 | 0.7919 |
| **NN** | image | 1000维VGG特征 | 0.6394 | 0.6312 | 0.6333 | 0.6323 |
| **NN** | text,image | 768维Bert嵌入，1000维VGG特征 | **0.8260** | 0.8159 | **0.8322** | **0.8240** |

注：NN表示神经网络

首先对比仅使用文本数据下的传统机器学习方法与深度学习方法，贝叶斯分类器与SVM分类器最高达到约68%左右的准确度，而使用Bert模型提取文本特征后再使用多层神经网络进行分类的深度学习方法获得80.42%的准确度，可见深度学习方法能够更加有效地提取文本特征并具备更强大的分类能力。

接下来讨论各种模型与文本特征提取的搭配问题。（1）朴素贝叶斯分类器本质是计算单词在各类文本中出现的条件概率，因此特征向量需要明确表达单词是否在文本中出现，因此贝叶斯分类器应搭配词袋（BOW）嵌入使用。（2）SVM分类器工作原理为寻找一个最合适的超平面来分割特征向量，使用TF-IDF嵌入取得了与贝叶斯分类器同级别的性能，但SVM在超高维向量空间上的训练速度很慢，而IF-IDF向量最大长度与BOW向量相同，所以SVM配合IF-IDF难以充分挖掘超长特征向量中的信息，在本案例实验中，最高使用了10000维的IF-IDF嵌入，且分类性能已略微超过使用34517维BOW嵌入的贝叶斯分类器。（3）SVM搭配Word2Vec嵌入时，分类性能较差，且Word2Vec特征越长，分类效果越差，这可能是由于：① 微博数据集的所有训练文本量不多，对于Word2Vec的训练是不充足的； ② Word2Vec向量不像BOW或TF-IDF是稀疏向量，当其长度过长时，不是线性可分的。（4）Bert模型是NLP领域较新的模型，其使用基于注意力机制的Transform结构，在超大规模的语料集上进行无监督的预训练，能够有效捕获文本的特征，搭配神经网络进行文本分类可以得到很好的效果。

最后讨论单模态与多模态的谣言检测。可以看到，同时使用文本与图像特征的多模态神经网络分类器取得了最佳的准确率82.60%。基于文本的单模态分类器准确率略低与多模态分类器，为80.42%；而基于图像的单模态分类器准确率较低，为63.94%。可以看出，文本对于多模态谣言检测任务极其重要，而图像虽然不如文本特征强大，但从63.94%的单模态分类结果来看，图像对谣言检测是有意义的，且加入后可以得到一定的性能提升。本案例使用在ImageNet上预训练的VGG网络进行图像特征提取，但ImageNet的图像可能与新闻中的图像差别较大，因此图像特征的提取有进一步的改进空间。

## 4 小结

该案例的选取基于多模态数据的社交平台谣言检测为主要问题，使用来源于微博平台的大量谣言与非谣言数据，结合数据挖掘、机器学习、深度学习的相关方法，进行谣言检测。该案例主题新颖，结合了社交媒体中信息可靠的现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习或深度学习方法完成任务，使用其它公开数据集进行实验研究，并考虑针对实际应用的进一步拓展。

## 附录

1. 本案例提供配套的PPT、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），Pytorch（深度学习算法库），Gensim（自然语言处理）。

4. 本案例参考文献如下：

[1] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 795-816.

[2] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[4] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

1. <https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view> [↑](#footnote-ref-0)