ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.17, No.8 March 2021
E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

机器学习在垃圾邮件过滤中的实现

冯军军,李力

(四川信息职业技术学院,四川 广元 628017)

摘要:随着通信技术的发展,垃圾邮件越来越多,对个人和中小企业危害也越来越大。该文介绍垃圾邮件识别使用的数据集以及特征提取方法,包括词袋模型和词汇表模型,然后介绍朴素贝叶斯、支持向量机、多层感知机、卷积神经网络和循环神经网络在垃圾邮件过滤的实现,经过对比发现多层感知机和卷积神经网络效果最好。

关键词:垃圾邮件;特征提取;NB;SVM;MLP;CNN;RNN

中图分类号:TP393 文献标识码:A

文章编号:1009-3044(2021)08-0154-02

开放科学(资源服务)标识码(OSID):

1 引言

垃圾邮件是指收件人拒绝接收或者不同意接收但是仍然收到的邮件[1],主要包含商业类、广告类、培训类、推广类、报价类等邮件。如2020年护网爆出的案例,就是攻击方通过信息收集获取防守方用户的邮箱,通过群发邮件,引诱用户更新钓鱼flash,从而实现权限获取进而内网渗透获取关键信息。为避免垃圾邮件影响人们的正常工作和生活,垃圾邮件检测技术随之产生。传统的垃圾邮件检测方法有关键词、黑白名单、校验码法等[2]。这些传统的垃圾邮件检测方法通常存在检测效果差、易被逃避等缺陷,为此本文引入机器学习方法解决这一问题。

垃圾邮件检测的关键在于识别哪些邮件是垃圾邮件,可以将该问题归结于机器学习中的文本分类任务。关于此类任务的数据通常是高维数据,本文选择常见的机器学习算法来进行垃圾邮件过滤的实现。

2 数据集和工具介绍

数据集采用开源的Enron-Spam[3]数据集,实现过程中,采用python语言。主要采用的模块为TFLearn,该模块可以直接在GitHub上面进行下载。它可以快速搭建实验环境,容易实现深度神经网络,内置神经网络层、正则化器、优化器等,支持多输入、多输出等模式。

3 特征提取

垃圾邮件的特征提取,朴素贝叶斯算法、支持向量机算法和多层感知机算法,采用的词袋模型,该模型利用单词构成的集合,如果一个单词在文档中出现不止一次,统计其出现的频数。邮件的特征提取,采用的sklearn.feature_extraction.text模块的CountVectorizer函数。利用该函数,将文本进行词袋处理,获取对应的特征名称,从而获取词袋数据,完成词袋化。该数据集在实现过程中,把邮件当成一个字符串处理。然后在字符串处理过程中,过滤掉空白符,例如回车符、换行符等。最后遍历全部邮件文件,加载数据。在词袋模型中,把邮件数据进行向量化,将正常邮件标记为0,垃圾邮件标记为1。

垃圾邮件的特征提取,卷积神经网络算法和循环神经网络算法采用的是词汇表模型。词汇表模型是在词袋模型的基础上,根据邮件内容生产的词汇表对原有句子按照单词逐个进行编码。通过VocabularyProcessor()函数,定义文本的最大长度、词频的最小值、分词函数等。在文本的最大长度初始化时,如果文本的长度大于最大长度,那么会截断文本,反之则用0填充。在词频最小值初始化时,如果出现次数小于最小词频则不会被收录到词表中。本文中,通过VocabularyProcessor函数对获取的ham和spam数据,进行处理,获取词汇表。

4 深度学习实现

本课题采用朴素贝叶斯算法、支持向量机算法、多层感知机算法、卷积神经网络算法和循环神经网络算法五种机器学习的算法,对邮件数据集进行识别。实现处理流程,如图1所示。如图1所示,在实现过程中,首先将数据样本根据算法要求进行特征提取(朴素贝叶斯算法、支持向量机算法、多层感知机算法实现用的词袋模型,卷积神经网络算法和循环神经网络算法实现用的词汇表模型);特征提取后把数据集随机划分为训练集和测试集,测试集的比例为40%;接着根据机器学习算法在训练集上进行训练,获取模型数据;最后根据模型数据,在训练集上进行预测,从而验证算法的预测效果。

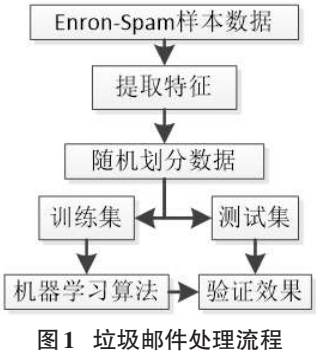

图1 垃圾邮件处理流程

4.1 朴素贝叶斯算法[4]

朴素贝叶斯算法(NB),该算法实现垃圾邮件分类的过程中,将数据集合随机分配训练集合和测试集合,实例化朴素贝

收稿日期:2021-01-22
基金项目:广元市院士工作站科研项目,项目名称:基于深度学习的云平台上垃圾邮件识别,项目号:2020KC09
作者简介:冯军军(1989—),男,讲师,硕士,主要研究方向:信息安全技术。

叶斯算法后，针对训练集进行训练，并针对测试集进行预测，最后输出评估结果的准确度和TP、FP、TN、FN（FN：False Negative，真实邮件判定为垃圾邮件，FP：False Positive，垃圾邮件判定为正常邮件，TN：True Negative，垃圾邮件判定为垃圾邮件，TP：True Positive，正常邮件判定为正常邮件）4个值。该方法中，由于最大特征数对结果有影响，把最大特征数，从1000到20000对评估准确度进行测试。发现最大特征数在13000左右的时候，系统准确率最大。但是随着特征数越大，耗时越大。根据结果，最终选择最大特征数为5000的情况下，整个系统准确度为94.33%，结果如表1所示。

测试关键代码如下：

```
gnb = GaussianNB()
gnb.fit(x_train,y_train)
y_pred=gnb.predict(x_test)
```

最后调用metrics.accuracy_score函数和metrics.confusion_matrix函数，输出TP、FP、TN、FN。

表1　朴素贝叶斯算法的验证结果

类型名称	相关	不相干
检索到	5937	632
未检索到	133	6875

4.2 支持向量机算法[5]

支持向量机算法(SVM)，该算法实现垃圾邮件分类过程中，数据特征提取、训练集获取数据模型、测试方法与朴素贝叶斯算法一致。最终在词袋最大特征数为5000的情况下，整个系统准确度为90.61%，其验证结果如表2所示。

测试代码如下：

```
clf = svm.SVC()
clf.fit(x_train, y_train)
y_pred = clf.predict(x_test)
```

最后调用metrics.accuracy_score函数和metrics.confusion_matrix函数，输出TP、FP、TN、FN。

表2　支持向量机算法的验证结果

类型名称	相关	不相干
检索到	5330	1239
未检索到	27	6891

4.3 多层感知机算法[6]

多层感知机算法(MLP)，该算法实现过程中，构造两层隐藏层，每层节点数分别为5和2。根据词袋模型，将数据集进行特征提取，按照分配的训练集和测试集，对于训练集进行多层感知机算法进行实例化，然后根据数据模型，对测试集进行预测，最后输出评估结果的准确度和TP、FP、TN、FN4个值。在词袋最大特征数为5000的情况下，整个系统准确度为98.01%，其验证结果如表3所示。代码实现，调用clf.fit()，传入训练集数据，然后调用clf.predict()，获取测试集的结果。最后调用metrics.accuracy_score函数和metrics.confusion_matrix函数，输出TP、FP、TN、FN。

表3　多层感知机算法的验证结果

类型名称	相关	不相干
检索到	6406	163
未检索到	105	6813

4.4 卷积神经网络算法[7]

卷积神经网络算法(CNN)，该算法实现过程中，特征提取采用词汇表模型，将数据集合随机分配训练集合和测试集合。实例化过程中，其中卷积神经网络模型，使用3个数量为128核，长度分别为3、4、5的一维卷积函数处理数据。使用卷积神经网络算法在训练集上训练，通过对训练数据进行了5轮训练，获取数据模型，使用模型数据在测试集上进行预测，最终实现对测试数据集的准确度为98.30%。代码实现中，调用tflearn.DNN()，根据网络，生成model对象。然后调用fit()，设置n_epoch为5，表示5轮训练，设置batch_size为100，表示一次用100个数据计算参数的更新。

4.5 循环神经网络算法[8]

循环神经网络算法(RNN)，该算法实现过程中，特征提取及数据处理与循环神经网络算法一样。在训练集实例化循环神经网络算法过程中，定义循环神经网络模型，使用最简单的单层LSTM结构。使用循环神经网络算法在训练集上训练，通过对训练数据进行了5轮训练，获取数据模型，使用模型数据在测试集上进行预测，最终实现对测试数据集的准确度为94.88%。代码实现中，调用lstm()函数，实现单层LSTM结构。调用tflearn.DNN()，根据网络，生成model对象。然后调用fit()，设置n_epoch为5，表示5轮训练，设置batch_size为10，表示一次用10个数据计算参数的更新。

5 结束语

以Enron-Spam数据集为训练和测试数据集，本文通过词袋模型和词汇表模型对数据集进行特征提取，通过NB、SVM、MLP、CNN和RNN，五种机器学习算法实现了垃圾邮件识别。通过比较发现，MLP和CNN的识别率很好，达到了98%以上。同样，在朴素贝叶斯算法实现的过程中，发现并非是词袋抽取的单词个数越多，垃圾邮件的识别率最高，而是有个中间点可以达到最大效果。总之，随着机器学习算法的发展，在垃圾邮件识别过程中，机器学习算法的应用会越来越多。

参考文献：

[1] 罗婧雯.垃圾邮件过滤技术综述[J].电脑知识与技术,2016,12(14):13-14.

[2] 李敬瑶.反垃圾邮件过滤技术方法的研究[J].福建电脑,2016,32(10):61-62.

[3] Enron-Spam数据集http://www2.aueb.gr/ users/ion/data/enron-spam/(DB/OL).

[4] 彭革.基于朴素贝叶斯算法在垃圾邮件过滤中的研究综述[J].电脑知识与技术,2020,16(14):244-245,247.

[5] 徐娟,卞良.基于SVM的中文垃圾邮件预测系统研究[J].数字技术与应用,2020,38(1):38-39.

[6] 赵俊生,候圣,王鑫宇,等.基于集成学习的图像垃圾邮件过滤方法[J].计算机工程与科学,2020,42(6):1049-1059.

[7] 马义超.基于卷积神经网络的手写数字识别算法研究与应用[D].焦作:河南理工大学,2019.

[8] 伍逸凡,朱龙娇,石俊萍.人工神经网络在信息过滤中的应用[J].吉首大学学报(自然科学版),2019,40(3):17-22.

【通联编辑：代影】