基于AdaBoost算法的情感分析研究

摘 要

随着互联网的快速发展，各类社交媒体平台如微信、QQ等也与日俱增，而微博更是集成了传统网站、论坛、博客等的优点，并加上了人与人之间的互动性、关系亲密程度等多种智能算法，并以简练的形式让数据爆发性的传播，促进了人与人之间的交流。网民可以通过微博来分享自己的生活，同时抒发自己的喜怒哀乐。因此对微博每天产生的信息量的分析和利用的需求显得更为迫切。

情感分析，也称倾向性分析、意见抽取和意见挖掘。主要是通过对带有情感色彩的主观性文本进行分析、处理、归纳然后进行推理的过程。而微博，人口基数大，涉及的话题广泛，对人们的日常生活产生了不可估量的影响，而对微博的情感分析，更是有着十分重要的意义。为此，本文针对了微博文本的情感分析进行了如下几个工作。

首先，使用微博官方的API对微博进行抓取，进行分类标注。然后，对微博文本进行预处理，主要包括去掉无意义，对微博文本没有影响的词语。其次，使用SVM算法对文本进行初步的筛选，主要是去除特别明显的广告等无关性的微博。最后使用朴素贝叶斯对微博进行情感分析，将微博分为积极、消极、客观三类，同时使用AdaBoost算法对朴素贝叶斯算法进行加强。

这些带有情感信息的微博是非常宝贵的资源，通过情感分析可以获取网民的此时的心情，对某个事件或事物的看法，可以挖掘其潜在的商业价值，还能对社会的稳定做出一定的贡献。

关键词：情感分析 AdaBoost 朴素贝叶斯 文本分类 数据挖掘

Research of Sentiment Analysis based on AdaBoost Algorithm

**ABSTRACT**

With the rapid development of Internet, various social medias like WeChat, QQ and Weibo are also growing with each passing day. Weibo which not only integrates the advantages of traditional websites, forums, blogs, but also leads into the introduction of a comprehensive interactive relationship between content, quality of intimacy and multiple algorithm promote the data communication explosively. And people can communicate more and more conveniently. Users can share their own lives by Weibo, and express their emotions. So, the need for analysis and use of information generated by Weibo every day is urgent.

Sentiment analysis is a process of analyzing, summarizing and reasoning subjective texts with emotion. And Weibo, based on a large population , involving a wide range of topics, has a great impact on people's daily life. And the emotional analysis of Weibo is a very important significance. In result, this paper focuses on the emotional analysis of Weibo text as follows.

First of all, micro-blog's official API is used for micro-blog crawl and then classifying weibo. After that, the Weibo text needs to be preprocessed by removing meaningless text, which does not affect the words. Secondly, the SVM algorithm is used for the preliminary screening of the text, mainly to remove the particularly obvious advertising and other unrelated with Weibo. Finally, using Bayes algorithm to analyze Weibo. Weibo text is divided into three types: positive, negative and objective, and AdaBoost algorithm is used to strengthen the Bayes algorithm.

The emotional information of Weibo is a very valuable resource. Through the analysis of Internet users, we can get emotional feelings, views of things or events at this time, also we can mine the potential commercial value and make a contribution for social stability.

**Key Words**: Sentiment analysis AdaBoost Naive-Bayes Text Categorization Data Mining

目 录

[前 言 1](#_Toc483924828)

[第一章 概述 2](#_Toc483924829)

[1.1 研究背景和现状 2](#_Toc483924830)

[1.2 情感分析的概念 2](#_Toc483924831)

[1.3 本文的内容安排 4](#_Toc483924832)

[第二章 微博的获取与清理 6](#_Toc483924833)

[2.1 概述 6](#_Toc483924834)

[2.2 微博的反爬虫机制 6](#_Toc483924835)

[2.2.1 通过Headers反爬虫 6](#_Toc483924836)

[2.2.2 基于用户行为的爬虫 6](#_Toc483924837)

[2.2.3 动态页面的反爬虫 7](#_Toc483924838)

[2.2.4 微博的反爬虫 7](#_Toc483924839)

[2.3 微博的获取 7](#_Toc483924840)

[2.3.1 微博API的获取 7](#_Toc483924841)

[2.3.2 模拟登录 8](#_Toc483924842)

[2.3.3 微博抓取与存储 9](#_Toc483924843)

[2.4 微博的分词与降噪 9](#_Toc483924844)

[2.4.1 概念 9](#_Toc483924845)

[2.4.2 分词 9](#_Toc483924846)

[2.4.3 删除URL 10](#_Toc483924847)

[2.4.4 删除用户名 11](#_Toc483924848)

[2.4.5 去除停用词 11](#_Toc483924849)

[2.5 本章小结 11](#_Toc483924850)

[第三章 SVM初步分类 12](#_Toc483924851)

[3.1 概念 12](#_Toc483924852)

[3.1.1 线性分类 12](#_Toc483924853)

[3.2 SVC 13](#_Toc483924854)

[3.2.1 线性支持向量分类机 13](#_Toc483924855)

[3.2.2 算法描述 14](#_Toc483924856)

[3.3 实验 15](#_Toc483924857)

[3.3.1 选取特征 15](#_Toc483924858)

[3.3.2 降噪 15](#_Toc483924859)

[3.3.3 特征降维 15](#_Toc483924860)

[3.3.4 将文本样本变成特征显示 16](#_Toc483924861)

[3.3.5 将文本随机分成训练集、测试集 16](#_Toc483924862)

[3.3.6 进行训练和预测 16](#_Toc483924863)

[3.4 本章小结 17](#_Toc483924864)

[第四章 利用贝叶斯定理进行情感分析 18](#_Toc483924865)

[4.1 引言 18](#_Toc483924866)

[4.2 贝叶斯定理 18](#_Toc483924867)

[4.2.1 高斯朴素贝叶斯 20](#_Toc483924868)

[4.2.2 伯努利贝叶斯 20](#_Toc483924869)

[4.2.3 多项式朴素贝叶斯定理 21](#_Toc483924870)

[4.3 本文中的多项式朴素贝叶斯 21](#_Toc483924871)

[4.3.1 算法过程 21](#_Toc483924872)

[4.3.2 拉普拉斯平滑 22](#_Toc483924873)

[4.4 实验 22](#_Toc483924874)

[4.4.1 分词 22](#_Toc483924875)

[4.4.2 特征提取 23](#_Toc483924876)

[4.4.3 向量化 23](#_Toc483924877)

[4.4.4 朴素贝叶斯分类 24](#_Toc483924878)

[4.4.5 测试 24](#_Toc483924879)

[4.4.6 计算准确率 25](#_Toc483924880)

[4.5 本章小结 25](#_Toc483924881)

[第五章 利用AdaBoost加强分类器 26](#_Toc483924882)

[5.1 集成学习 26](#_Toc483924883)

[5.1.1 发展历史 26](#_Toc483924884)

[5.1.2 Bagging方法 26](#_Toc483924885)

[5.1.3 Boosting方法 26](#_Toc483924886)

[5.2 ADABOOST 27](#_Toc483924887)

[5.2.1 概念 27](#_Toc483924888)

[5.2.2 举例 27](#_Toc483924889)

[5.2.3 算法描述 29](#_Toc483924890)

[5.3 ADABOOST提升朴素贝叶斯文本分类 32](#_Toc483924891)

[5.4 多类问题 33](#_Toc483924892)

[5.4.1 AdaBoost.M1算法 33](#_Toc483924893)

[5.4.2 AdaBoost.SAMME 34](#_Toc483924894)

[5.4.3 AdaBoost.SAMME.R 34](#_Toc483924895)

[5.5 实验 35](#_Toc483924896)

[5.5.1 二分类 36](#_Toc483924897)

[5.5.2 多分类 37](#_Toc483924898)

[5.6 本章小结 39](#_Toc483924899)

[第六章 总结与展望 40](#_Toc483924900)

[6.1 本文主要内容总结 40](#_Toc483924901)

[6.2 存在问题及未来研究展望 40](#_Toc483924902)

[参考文献 42](#_Toc483924903)

[致 谢 44](#_Toc483924904)

# 前 言

随着互联网的快速发展，各类社交平台越来越流行，甚至开始占据人们的大量时间，每天产生的信息量也与日俱增，对其分析与利用也迫在眉睫。而微博，不仅给予了用户更自由、更快捷的方式来进行沟通、表达观点、记录心情等，逐渐发展成了国内最热门的社交平台之一。根据去年的微博财报显示，自微博上市以来，其月活跃用户数量已经连续了11个季度保持了30%以上的同比增长，2016年，微博活跃用户数量整年净增达到了7700万，达3.13亿，同时，日活跃用户数也增长到1.39亿[1]。由此可见，微博的强大影响力已经深深的吸引了更多的人加入。而对微博的情感分析，不仅可以获取网民的此时的心情，对某个事件或事物的看法，还可以获取其潜在的商业价值，还能对社会的稳定做出一定的贡献。

情感分析（Sentimentanalysis），也称意见挖掘(Opinion Mining)，主要是对带有感情色彩的主观性文本进行分析、处理、归纳然后进行推理的过程，例如对产品，话题，政策的意见[2]。利用这些分析的结果，消费者可以深入了解商品的实用性，从而优化购买的决策，同时，生产者和经销商可以改善自己的服务，从而赢得竞争的优势。随着信息时代的到来，越来越多的公司开始组建数据分析团队对自身公司的数据进行挖掘、分析。比如某服装公司想调查自己制作的服装的受喜爱程度，就可以从服装的评论入手，挖据文本内容，判断留下评论的用户对服装的喜好态度，积极的、消极的或者是中性的评价。

情感分析作为Web挖掘中新兴的一个领域，对其不同角度的研究也越来越多，比如识别商品评论的信息、判断客户的褒贬态度等。姚天昉[3]等人对情感分析的研究现状做了如下总结：①首先，介绍情感分析的定义和研究的目的；②从主题的识别、意见持有者的识别、情感描述的选择和情感的分析四个方面进行评述，并介绍了一些成型的系统；③讨论中文情感分析的研究现状。而本文将要从微博的符号、词语等粒度上，用情感分析的方法对微博文本进行分类。

虽然研究者在文本挖掘展开了越来越多的研究，但是对各个领域的深入挖掘依然处在探索阶段。而微博，作为一个越来越吸引用户的社交平台，涉及的内容十分广泛，如娱乐、影视、体育等，不同内容针对不同的领域都有着不可忽视的影响。本文主要结合文本情感分析领域的研究结果以及现有的微博情感分析方法，将对微博的情感分析分为四大类：文本预处理、SVM过滤无关信息、进行情感分类、加强分类算法。

# 

# 第一章 概述

## 1.1 研究背景和现状

随着互联网的快速发展，互联网在平民百姓中显得极其重要，人们从只能口述交流、到写信、再到现在电话语音交流，而互联网应用集合了以前交流方式，提供了更加快捷方便的功能让人们的交流越来越方便。微博是当下热门的互联网应用之一，其用户日趋增长，每天微博产生的信息量也越来越大，不仅仅在热点事件中有着不可估量的影响力，而且已经深入了网民的生活中，成为了用户不可缺少的一部分。于是，对微博的文本挖掘研究产生的价值也随之上升。

情感分析，正成为各界关注越来越关注的领域，主要用来识别一段文字的情感倾向。现实中，文本中能够看出人们表达出来的情感是十分复杂的，目前的自然语言处理的研究一般将倾向性划为正向和负向，这种研究方法使得情感分析与机器学习结合相当简便。通常情况下，对某些文本进行情感标注，之后划分为训练集和测试集，使用机器学习方法来进行分类，然后优化算法，最终得到分类结果。

## 1.2 情感分析的概念

情感分析，顾名思义，又被称为倾向性分析和意见挖掘，通常使用带有情感色彩的词语对文本进行分析、处理、归纳、推理等过程[4]，如：从购物网站上分析用户对某一件商品是好是坏的过程，从电影评论网站上分析用户对某部电影的评价，从音乐平台上的评论来鉴赏网民对某首音乐的喜爱程度等。其中，斯坦福通过公共平台如Twitter分析网民在2008~2009年金融危机的心情[3]，如图1.1。

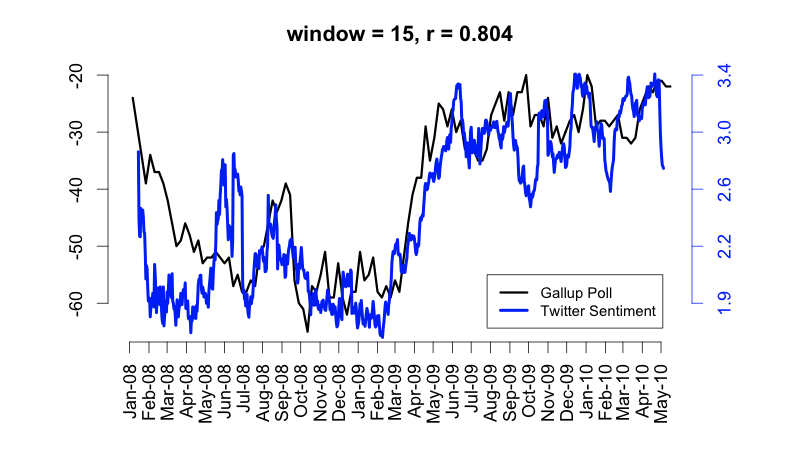


图1.1 Twitter用户的情感

同时，也有一些网民使用Twitter来分析了各个州的网民对2016年特朗普和希拉里的美国总统选举的支持[6]。

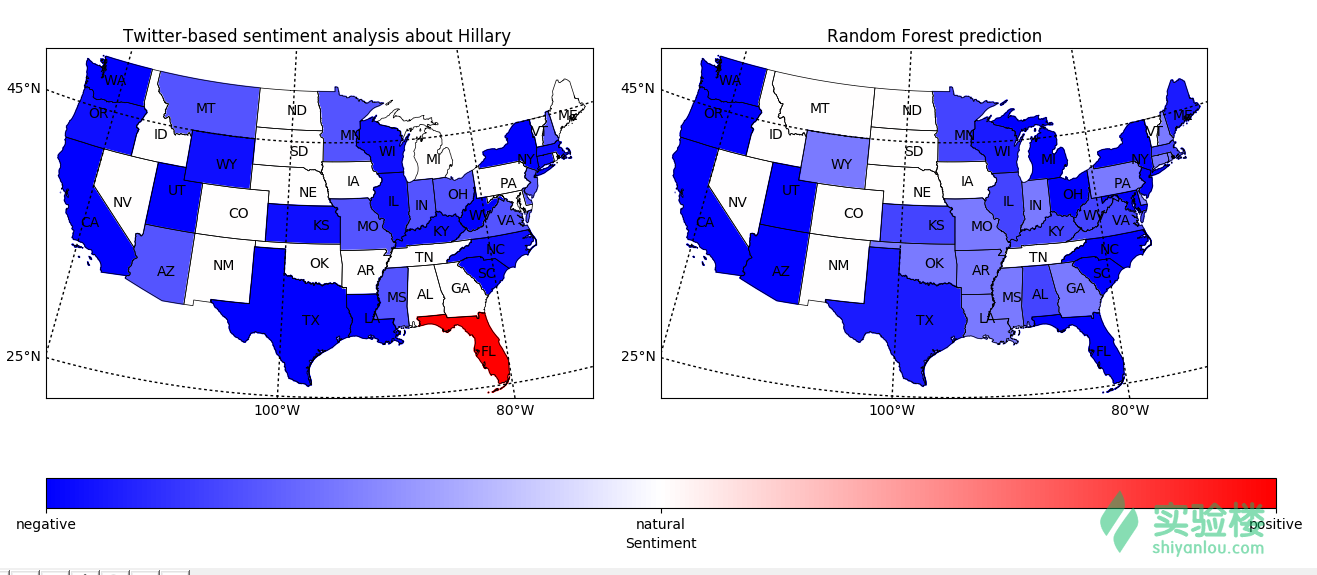


图1.2 每个州对希拉里的情感

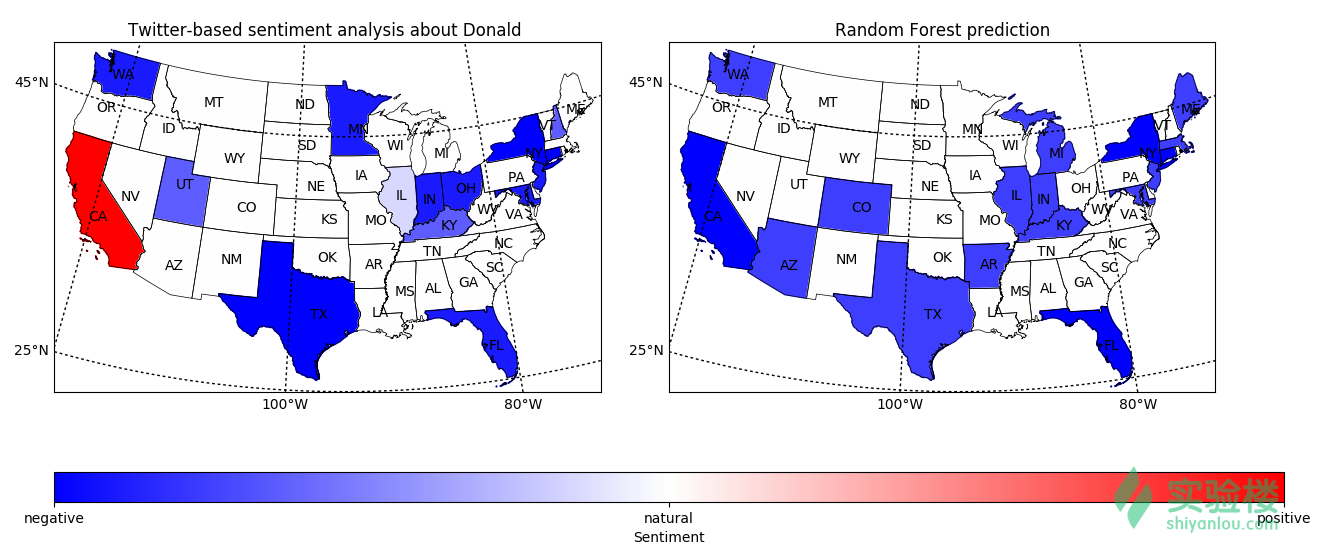


图1.3 每个州对川普的情感

当下，情感分析的任务一般采用机器学习来进行分类。例如在一般购物网站中，一般都有商品的评论，因此使用其作为机器学习标注后的文本，之后使用机器学习方法来分类，最终构建一个情感分类器来对其他评论进行预测。但是对于微博这种数量庞大的互联网文本，想要对大量的微博进行标注是几乎不可能的，只能使用少量的人工标注的微博文本来进行机器学习。

目前，机器学习对情感分析的分类受制约的还是多种情绪的表达，以及网络新兴的流行的词汇等。由此可见，构建情感词典显得特别重要，在微博预料中需要全面高效的捕捉情感的基本单元，才能准确的计算出每条微博的情感倾向。使用机器学习进行情感分析的一般流程如下图1.4所示。



图1.4 情感分析一般流程

## 1.3 本文的内容安排

论文全文分为6章：

第一章主要是介绍研究的背景和研究现状、情感分析的概念等研究工作，指出这些工作在当前问题下存在的不足，并基于此提出该研究的关键问题。

第二章对微博的获取与清理，讲解了如何通过微博API的获取Token，然后模拟登录获取最新微博，最后介绍了微博的文本预处理，主要去除掉了对微博分类有影响的字符和无关信息，为开展研究提供支持。

第三章介绍了使用svm对微博的初步分类，主要分为无关类和有关类，无关类主要包括广告、无关字符（如韩文日文等）、重复字段（如分享图片）等。

第四章针对微博进行标注，然后选取特征词，在使用贝叶斯分类，将微博分为积极、消极、中性三个分类。

第五章讲解了如何使用了AdaBoost来加强贝叶斯分类器的分类效果，同时对AdaBoost二分类和多分类的问题进行了讨论。

第六章总结本论文的研究内容，并对可能的研究方向进行展望。

# 第二章 微博的获取与清理

## 2.1 概述

从上一章的研究背景所述，微博已成为社交平台的十分突出的新媒体。该平台以特定的方式授权开发者获取微博的部分数据，以此让开发者分析、发布、处理等，并且能够让开发者深入研究和商业运作。

针对微博数据的抓取和存储，目前主要有两种方式：

（1）根据微博官方提供的API接口。这些接口基于REST实现的HTTP协议，以JSON结构化的方式作出响应。但是，这种接口需要一定的权限，还有接口的请求频次限制次数，甚至对于接口的限制的速率限制。

（2）基于微博网页的解析。这种方式需要解析网页，如果网页代码有改动，响应的抓取方式也要有改变，对于大量的抓取，需要破解微博的反爬虫机制（使用代理ip，不同账号等）。

## 2.2 微博的反爬虫机制

防止爬虫一般从三个方面入手：分析网页请求的headers，监督用户访问网站的行为，调整网站中的目录和数据加载的方式。前面两种比较常见，大部分网站都是从这两个角度来反爬虫。第三种会应用一些ajax来反爬虫。

### 2.2.1 通过Headers反爬虫

目前，一般网站都会检测网页请求中Headers的User-Agent，有的甚至还要检查网站的Referer。如果遇到这类反爬虫机制，我们可以直接在代码中添加Headers和Referer以此来绕过检查。对于这些网站，在代码中添加或修改其中的Headers和Refer就能很好的绕过。

### 2.2.2 基于用户行为的爬虫

用户访问网站的行为也是目前主流网站常用的检测手段，比如：同一IP在短时间内多次访问了同一个页面，还有的事同一个账户短时间内多次进行相同操作。对于这种情况，我们可以使用IP代理来解决。现在网上有收费的和免费的IP代理，我们可以爬去这些IP代理存储起来，然后每请求几次就更换一次IP。

### 2.2.3 动态页面的反爬虫

还有一部分网站，数据是通过ajax或者js请求生成的。我们可以使用浏览器对访问网站中的请求的进行分析。如果能找到ajax请求，分析其含义后可以使用上面两种方法解决，获取对应的数据。

如果不能获取ajax的请求，可以调用selenium+phantomjs框架，调用其浏览器内核，来模拟人为操作以及触发页面的js脚本。

### 2.2.4 微博的反爬虫

微博中的反爬虫使用了以上三种机制，验证客户端的Headers，同时对访问量多大的同意IP进行禁止访问，使用Ajax进行数据传输。要想破解此类爬虫，必须使用IP代理，同一账号不同时间访问，添加Headers等。微博中的Headers添加如下：

conn.request('post', '/oauth2/authorize', postdata,  
 {'referer': self.\_author\_url,  
 'content-type': 'application/x-www-form-urlencoded',  
 'user-agent': 'mozilla/5.0 (windows nt 10.0; wow64) applewebkit/537.36 (khtml, like gecko) chrome/50.0.2661.102 safari/537.36',  
 'cookie': 'your cookie'})

本实验中，主要是基于微博开放平台提供API，然后申请一个statuses/public\_timeline接口，获取到最新的公共微博。

## 2.3 微博的获取

### 2.3.1 微博API的获取

使用新浪登录之后，在新浪微博开放平台申请一个应用，由于本文需要用到官方的sdk，所以需要应用中的key、secret和回调页



图2.1 申请微博应用

回调页如下：



图2.2 回调页

App Key和App Secrete如下：



图2.3 App Key和App Secrete

### 2.3.2 模拟登录

由于微博中的接口需要获取权限，所以需要手动截取token，比如链接中：<https://api.weibo.com/oauth2/authorize?code=7dded6d1b81bdc341cc75d585b566492>

链接的token就是code=7dded6d1b81bdc341cc75d585b566492。本文将采用模拟登陆微博并获取token后直接调用接口的方法，省去了手动输入的麻烦。

### 2.3.3 微博抓取与存储

使用官方的python sdk，然后使用statuses/public\_timeline的接口一次性获取200条微博。其连接代码如下：

json\_str = client.get('statuses/public\_timeline', uid=uuid, separators=(',', ':'), count=200)

由于服务器返回的格式是json的，所以将json中所需要的key和value存储入数据库即可。

## 2.4 微博的分词与降噪

### 2.4.1 概念

微博作为一个社交平台，不仅拥有能够迅速的传播特点，并且成为了商家用来推广产品的重点发布平台。在微博中，大量广告、营销类账号的出现，给微博的情感分析造成很大的困难。所以，针对微博的文本预处理变得极其重要。

### 2.4.2 分词

中文的分词是将一个汉字序列变成一个个单独的有意义的词汇。文本挖掘首先要以中文分词为前提。目前常用的中文分词软件主要有以下几种：

（1）BosonNLP：玻森的中文语义开放平台，主要提供了一个方便、性能强大的中文自然语言分析的平台。

（2）NLPIR：是中国科学院计算机研究所的一个产品，积累了多年研究工作，暂时是目前世界上最好的中文分词工具。

（3）结巴分词：一款开源在GitHub上面的中文分词工具，提供了python、java等多语言的接口，而且能够识别繁体字，立志成为最好的中文分词工具。

（4）IKAnalyzer：一款开源的java分词工具，最初是以lucene为应用主体的，之后结合了词典分词和语法分析算法的分词组件。

本文中将采用中科院的NLPIR和开源的JIEBA进行分词，其中NLPIR的分类例子如下：

不让我上桌吃饭我还不会自己抢吗！[doge][doge][doge]（投稿：@还没怀上的葛一他麻麻）http://t.cn/RqKTebK

如果不降噪进行分词：

['（', '不', '@还没怀上的葛一他麻麻', '投稿', '抢', '！', '会', '自己', '还', 'RqKTebK', '桌', '上', '吃饭', '[doge]', '让', '：', 'http://t.cn/', '）', '吗', '我']

### 2.4.3 删除URL

垃圾微博中的内容一般较短，而且一般文字后面都带有链接，由此才能将用户导向网页的入口，如下面几条：

【领 10 元 优惠 券】【券 后价 19.9元】【包邮】 粉丝福利购：http://t.cn/R6j6YyX ​"

"申马 帆布鞋 =26.8 领券拍：http://t.cn/R6lLnsV http://t.cn/R6n9kRO ​"

"玩游戏？找优惠？上莫莉幻想！http://t.cn/RxmHa1i ​"

"蓝牙耳机 =49 领券拍：http://t.cn/R6nIheq http://t.cn/R6nI2Fp ​"

"赫拉 气垫bb =58 领券拍：http://t.cn/R6CjMuM http://t.cn/R6QsX8f ​"

由上述内容可以知道，一般情况下的普通URL链接都是较长的字符串，如果保持原链接会占用微博的资源。因此，微博希望能够将原本的“长链接”变成缩短的短链接。微博中使用散列（hash）索引，将原始链接对应成一个较短的、一一对应的字幕、数字串组合。如：

<http://t.cn/R6QsX8f>

其原链接为：

<http://shop.sc.weibo.com/h5/goods/index?iid=110054105972300003317973>

URL中带有有用的信息很少，一般都是广告的导向和用户的定位。我们使用SQL从数据库中查找含有URL的微博数量统计。

表2.1 统计结果

|  |  |  |
| --- | --- | --- |
| 统计项 | | 结果 |
| 含有URL的微博数量 | 1756（总量：4780） | |
| 平均引用次数 | 2.72 | |

由此可见，URL在微博中的引用量是特别大的，在情感分析前，要对微博文本进行适当的清理，从而去除这些无用的URL，降低这些URL对情感分析的影响。

### 2.4.4 删除用户名

微博文本中的用户名一般用来提醒该用户，但是，大部分微博用户的用户名毫无规律性，如：@real\_\_pcyyyyy、@CloverH静、@baekhyunee7永远像25岁一样年轻等，对分词器来说有较大的影响，比如JIEBA分词会把“@baekhyunee7永远像25岁一样年轻”拆分成['@', 'baekhyunee7', '永远', '像', '25', '岁', '一样', '年轻']，其中“年轻”会对用来构建的词性特征造成影响，所以，对于用户名的出去也是极其重要的。

### 2.4.5 去除停用词

停用词（STOP WORDS），在词典中的意思为：对文本中表达的意义并不起什么作用的词语。在SEO中，为了节省存储空间和提高检索速度，搜索引擎会在搜索时自动忽略某些字或词，这些字或词便是停用词。

停用词在一定程度上相当于过滤词，但是过滤词的范围比较大，通常包含色情、政治、暴力等敏感信息，停用词则没有这个限制。通常情况下，停用词可分为两类：

（1）使用广泛，过于频繁的一些单词。比如“我”、“你”之类的词几乎在很多文档中都会出现，对于搜索引擎来说，这类词无法保证准确的搜索结果，还会降低效率。

（2）文档中使用频率很高，但是实际意义不大的词。主要包括语气助词、副词、介词、连词等，在文本表达中没有变现出任何意义。为了增加情感分析的准确性，我们需要对微博文本中去除停用词。

## 2.5 本章小结

近年来，大数据、数据挖掘、机器学习和可视化的流行是始终无法绕开数据清洗这个环节，几乎所有的数据科学都来源于这些乱七八糟的数据上，而本文面对的主要是微博文本的分析。众所周知，随着微博的日趋增长，其中的流行词越来越没有规则，如特殊的符号、毫无逻辑的网名、甚至是看似没有关联的句子等。本章主要讲解了如何获取最新微博，以及对微博的初步降噪，去除URL、删除用户名，去除停用词等，为下一步的初步分类提供了基础。

# **第三章 SVM初步分类**

## 3.1 概念

SVM理论由V.Vapnik提出，不仅用于模式分类还可以用于非线性回归，主要做法是在样本点所在的向量空间中，找出一个使分类要求的最优分类超平面，该平面能把不同类的样本分开，不仅能够满足分类精通的要求，还能使两侧的空白区域（分类间隔）最大化[7]。

支持向量机（support vector machine，SVM）现在已经越来越被研究者关注的一种分类算法。该分类算法具有坚实的统计学理论基础，而且还在许多实际应用比如手写数字的识别、文本分类等都展示了特别明显的实践效用。

原理简介：SVM是通过一个非线性映射p，把样本空间映射到一个高维乃至无穷维的特征空间中，使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题[8]。常用的SVM主要依赖于不同的核函数，常用的几种核函数如下：

1. 高斯径向基核：； （3.1）
2. 多项式核： （3.2）
3. B-样条核：； （3.3）
4. 傅里叶核：； （3.4）

### 3.1.1 线性分类

先举个简单的例子，一个以圆点和“x”的点分布在二维平面上，用一条直线的线把圆点和“x”的点分开来。这条直线的线也就是我们所说的超平面，假设超平面两边的数据点对应的是-1和1。

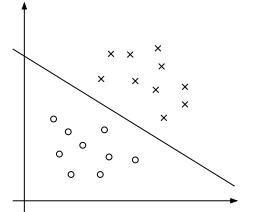


图3.2 线性SVM

接着，我们可以令分类函数：

（3.5）

显然，如果，那么x是位于超平面上的点。我们不妨要求对于所有满足的点，其对应的y等于-1 ，而则对应y=1的数据点。

SVM分类算法主要包括分类和回归分析两个方面。本文主要为文本分类即分类问题，所以采用线性支持向量分类机（SVC）展开。

## 3.2 SVC

支持向量分类机（SVC）最早用在了在手写数字的识别上。同时，Vapnik和他的团队最初是在处理两类问题的基础上，提出了处理多类问题的支持向量机。

Joachims首先使用支持向量机与文本分类上。文本分类的应用很广，主要包括垃圾邮件过滤、网页检索、情感分析等。本文主要通过SVC来对文本进行初步分类，主要是分出无关类的微博，方便下一步的情感分析。

### 3.2.1 线性支持向量分类机

上一节所述内容均是理想下的线性可分情况，但是现实中这种情况很少，甚至不存在。一般情况都是如图3.2所示中的非线性情况。于是，提出了线性支持向量分类机（LinearSVC）。线性支持向量分类器一般用在线性不可分的数据集，即将数据集当做近似可分。我们也可以通过求解得出对应的凸二次规划问题，同时求出超平面。



图3.2 非线性

本文中使用python sklearn库中的LinearSVC来对文本进行分类，其类似于SVC与线性内核参数，但是，它的实现方式是基于liblinear而不是libsvm，所以，它能有更加灵活的选择处罚和损失函数。其算法描述如下一节。

### 3.2.2 算法描述

直观上，我们假设存在一个最优的超平面，其公式为：

（3.6）

对于所有的点集x，都存在能够平行于最优超平面的边界面。即或者。如果要对图3.2中的进行划分，则需要放弃间隔值取正。本文中引入松弛条件，i=1,2,…,l，此时，问题转化为：。将此公式写成等价形式：

（3.7）

其中C>0为惩罚函数，利用拉格朗日变换，函数变为：

（3.8）

对，求导置0后代入原公式，得到起对偶：

（3.9）

对于LinearSVC的分类问题，就是寻找使得拉格朗日函数L最大，即：

（3.10）

其算法描述如下：

1. 设已知训练集：
2. 适当地选择一个惩罚函数C>0，构造并求解最优化问题：

（3.11）

使得:

（3.12）

对其求解后，我们可以得到最优方案。

1. 计算，选择的一个小于C的正分量，并根据此计算：

（3.13）

1. 之后构建一个超平面，来获取决策函数：

（3.14）

或者：

（3.15）

此算法同时也称线性软间隔分类机，由算法中可以看出，从理论上来说，能够确保唯一，但是不一定唯一。在现实的使用中，大部分只出现满足唯一性的充分条件，即该算法具有可操作性。

## 3.3 实验

### 3.3.1 选取特征

选取特征就是选出分类对象中所展现的部分特点，用来实现分类的依据。举个例子：

能拿奖品这种事一定要和大家说一说的，否则以后还怎么在一起愉快的玩耍[酷]我正在领取#微博等级专享礼#，这里的奖品太给力了，快去试试人品吧<http://t.cn/RLAxtki>。

特征是从大量的文本中选举出来的高频词，假设样本中有许多类似这样的句子，那么由上面该句子我们可以知道，“奖品”、“微博等级”是该句子的高频词，也是影响其分类的主要特征。

### 3.3.2 降噪

在上面这个例子中，经过降噪处理后：

['拿', '这种', '否则', '愉快', '我', '说', '力', '以后', '大家', '太', '博', '一起', '[酷]', '在', '还', '奖品', '了', '给', '事', '吧', '礼', '#', '这里', '能', '等级', '专', '要', '人品', '试试', '怎么', '正', '微', '享', '的', '领取', '一', '和', '去', '，', '快', '一定', '玩耍']

### 3.3.3 特征降维

特征降维,即降低特征的维数，降低复杂度。在自然语言处理中，多余的特征有时候会影响或误导学习器，特征太多就需要调整更多的参数，过拟合的风险也就越大。本文中，主要是把无关或冗余的特征删掉，特征数量减少了之后不经能够可以加快算法计算的速度，还可以减少干扰，提高分类的准确率。在自然语言处理中，主要是通过一定的统计方法来找出信息量丰富的特征。其中，这些方法包括：信息熵、卡方统计、词频、文档频率、互信息等。

本文中，使用的是词频统计，也就是选在语料库中出现频率高的词。比如我可以选择语料库中词频最高的n个词（降噪后）作为特征。之后，再将所有的原始文本转化为特征表示的形式。

### 3.3.4 将文本样本变成特征显示

所以，上诉文本经过选取特征之后如下：

[{'试试': 'True', '正': 'True', '礼': 'True', '人品': 'True', '专': 'True', '力': 'True', '事': 'True', '玩耍': 'True', '说': 'True', '博': 'True', '微': 'True', '[酷]': 'True', '享': 'True', '奖品': 'True', '领取': 'True', '太': 'True', '等级': 'True', '愉快': 'True'}, 'adv']

### 3.3.5 将文本随机分成训练集、测试集

在机器学习分类中，必须要有一定的数据用来训练，一定的数据用来测试，这样才能获取该分类器的准确度，之后可以根据实际情况构建出一个高效的分类器。

本文使用python中的random里的shuffle来将数据随机化，之后再将文本分为训练集和测试集。

### 3.3.6 进行训练和预测

我们使用python scikitlearn中的LinearSVC进行训练和预测，然后进行训练和预测，对实验中进行二十次迭代，得出结果绘制成图表如图3.3所示。

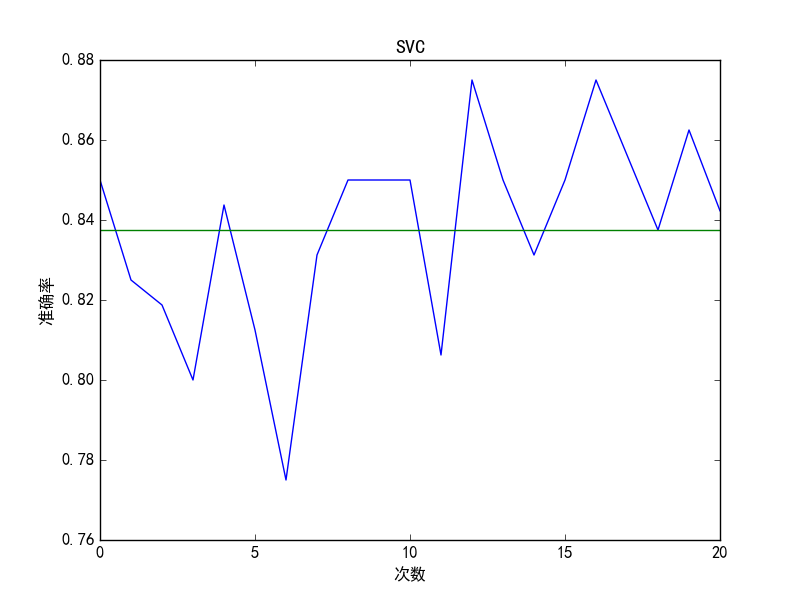


图3.3 实验结果

## 3.4 本章小结

本章讲解了使用Linear-SVC算法对微博进行初步的分类，去除掉一些特别明显的无关性的微博，比如：特别明显的广告、重复特别多的词语等。微博信息量的与日俱增，但是利用其来进行商业行为的更是到处遍布着整个微博，对于大数据量情感分析来说，这些无关性的微博很明显会对情感分析的性能产生干扰，影响的不仅仅是速度。所以，对微博进行初步分类，减少无关性的微博能够有效的提升情感分析的速度与效率。

# 第四章 利用贝叶斯定理进行情感分析

## 4.1 引言

贝叶斯方法作为一个历史悠久，有着坚实的理论基础的机器学习方法，不仅能够在同时处理很多问题时直接而又高效，而且很多高级自然语言处理模型也能够从它演化而来。贝叶斯方法，是研究自然语言处理问题的一个极其优秀的切入口。其流程主要如下图4.1。



图4.1 朴素贝叶斯情感分析基本流程

第一阶段——准备工作阶段。此阶段主要是对文本进行预处理，先对样本进行标注，之后根据词频筛选部分的特征词。该阶段输入的事所有待分类的样本，然后得出特征属性和训练样本。朴素贝叶斯的分类器的准确性主要由筛选出来的特征属性来决定。

第二阶段——分类器训练阶段。根据样本中的频率，然后由每个特征计算出每个类别的先验概率。此阶段主要是根据公式的机械计算。此阶段是朴素贝叶斯分类最重要的一个环节。

第三阶段——应用阶段。该阶段主要将测试样本进行输入，然后由分类器来计算出分类的记过。

## 4.2 贝叶斯定理

贝叶斯公式就一行：

（4.1）

而它其实是由以下的联合概率公式推导出来：

P(Y,X)=P(Y|X)P(X)=P(X|Y)P(Y) （4.2）

其中P(Y)为先验概率，叫做后验概率，P(Y,X)一般称作联合概率。

算法描述：

标准的朴素贝叶斯分类算法的执行过程如图4.2所示。

（1）对样本进行人工标记

（2）对不同类别的样本进行中文分词

（3）去除样本中的垃圾词条

（4）将整理后的词条做成特征组，分析并计算词条频率信息

（5）根据词条的信息，计算其先验概率。

（6）读取训练的样本，进行分词，降噪，然后形成样本特征组

（7）将测试样本的先验概率代入朴素贝叶斯公式并计算后验概率，得到最大概率的所属类别就是文本的类别

图4.2 朴素贝叶斯文本分类流程

本文中：将不同的经过第三章的提取（去掉无关类的微博）后，将剩下的微博分为三类：POS（积极）、NEG（消极）、Neural（客观）。

（1）一般情况下，积极即带有取向上、努力等思想和表现，如：屋里tyy生日快乐[噢耶][噢耶][噢耶]

（2）而消极则代表了负面的，不思进取的，失落等情感，如：有时候觉得人活着挺难的，不想说话[闭嘴]

（3）而客观的就表示仅仅陈述，不含有情感色彩的，如：一百本可以充实自己的外国图书，书荒的Mark

对于文本分类，常用的朴素贝叶斯主要存在三种不同的贝叶斯模型：高斯模型、多变量的伯努利模型和多项式模型

（1）高斯模型——它假设特征是正态分布的。它的一般使用场景是，给出人物的宽度和高度，判断这个人的性别。而情感分析，从给定推文文本中提取出词语的个数，不适合正态分布。

（2）伯努利模型——和高斯模型相似，更适于判断词语是否出现二值特征，而不是词频统计。

（3）多项式模型——它假设特征就是出现次数。这和我们是相关的，因为我们会把推文中的词频当做特征。

### 4.2.1 高斯朴素贝叶斯

某些特征很可能是连续型变量，比如说物体的长度和人的身高，这些特征可以转换成离散型的值。比如我们将人的身高进行划分，如表4.1。同时，我们也可以将身高用三个特征表示f1，f2，f3，结果如表4.2所示。

表4.1 身高特征划分方法一 表4.2 身高特征划分方法二

|  |  |
| --- | --- |
| 0~160cm | 1 |
| 160~170cm | 2 |
| 170+cm | 3 |

|  |  |
| --- | --- |
| 身高 | f1 f2 f3 |
| 0~160cm | 1 0 0 |
| 160~170cm | 0 1 0 |
| 170+cm | 0 0 1 |

不过这些方式都不够细腻，高斯模型可以解决此类问题。高斯模型假设这些一个特征的所有属于某个类别的观测值符合高斯分布。

### 4.2.2 伯努利贝叶斯

伯努利贝叶斯（Bernoulli Naïve Bayes，BNB）是最早基于朴素贝叶斯模型对文本进行分类的算法。模型中，一篇文本会被表示成欧式空间中的一个二进制变量，即，如果文本中在指示变量中出现的话，则将改值标为1，否则为0.给定测试文本d=<w1,w2···wm>，如果属性条件独立，则BNB根据下面的公式选出最大的后验概率来对文本进行分类：

c(d)= （4.3）

从之前别人的研究可以表明伯努利贝叶斯在数据量比较少的时候性能较好，但是，当数据量大的时候，性能远远比不上多项式模型。

### 4.2.3 多项式朴素贝叶斯定理

多项式朴素贝叶斯（Multinomial Naïve Bayes,MNB）是为了改进多变量伯努利朴素贝叶斯而提出的。该模型中，文本的表示是欧式空间中的一个带有单词频率信息的向量，当一个词语在一篇文章中出现时，该文本的对应的词语频率为1，否则为0。

在条件独立的情况下，给定测试文档d=<,,…,>,多项式朴素贝叶斯根据下面的极大后验概率来对文本进行分类：

（4.4）

## 4.3 本文中的多项式朴素贝叶斯

### 4.3.1 算法过程

训练集TrainingSet={(,), (,)…(,)}包含N条训练数据，其中， T是M维向量，属于K类中的一类。

首先，我们来计算先验概率p(y=)

（4.5）

其中I(x)为指示函数，若括号内成立，则计1，否则为0。

接下来计算分子中的条件概率，设M维特征的第j维有L个取值，则某维特征的某个取值，在给定某分类下的条件概率为：

（4.6）

通过学到的概率，给定未分类新实例X，就可以通过上述概率进行计算，得到该实例属于各类的后验概率p(y=|X)，因为对所有的类来说，分母的值都相同，所以只计算分子部分即可，具体步骤如下：

计算该实例属于类的概率。

（4.7）

确定该实例所属的分类y。

（4.8）

其中，最大的便是我们的分类。

### 4.3.2 拉普拉斯平滑

在计算先验概率和后验概率的时候，从样本中算出的概率值很有为0，会导致相乘的结果为0，这显然是不合理的，因为不能因为一个事件没有检测到就判断改事件的概率为0。我们可以通过分子和分母都分别加入一个平滑因子aa，就可以避免这个问题。更新过后的先验概率公式变为：

（4.9）

K表示类别，公式4.6变为如下：

（4.10）

是第j维特征的最大取值

可以证明，改进以后的仍然是概率。平滑因子kk=0即为实现的最大似然估计，这时会出现在本节开始时提到的0概率问题；而kk=1则避免了0概率问题，这种方法被称为拉普拉斯平滑。

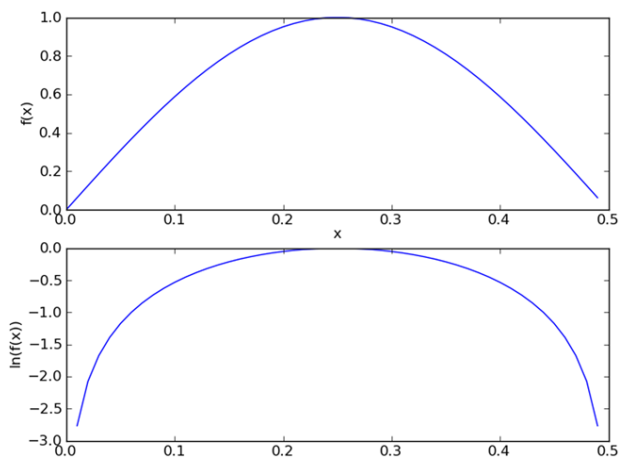


图4.2 拉普拉斯平滑曲线

## 4.4 实验

### 4.4.1 分词

结巴分词，是Github中开源的中文分词组件，该分词器能够支持三种分词模式：

1. 精确模式：能够将句子用最精确的方法切开，普遍适用于文本分析；
2. 全模式：能够把文本中所有成词的词语都扫描出来，虽然速度非常快，但是不能解决歧义。
3. 搜索引擎模式：在精确模式的基础上，会对长词再次切分，提高召回率，一般适用于搜索引擎分词。

本实验是基于标注的情感分析，故使用结巴分词的精确模式来进行分词。

### 4.4.2 特征提取

跟第三章中所述的一样，使用卡方统计筛选出频率最高的词。得出的结果使用词云如图4.3所示：



图4.3 词云

### 4.4.3 向量化

假如上述中的特征提取中构建出来的单词特征为[‘喜欢’,’失望’,’快乐’,’越来越好’,’晚安’]，长度为m，矢量化的时候如果一条微博为：生日快乐，晚安。那么，构建出来的矩阵为：[0,0,1,0,1]。

说明：

1.如果为n条微博，则构建出来则是n\*m的矩阵。

2.如果一条微博的某个特征出现次数多于一次，则进行累加，如，快乐快乐，矢量化之后变成：[0,0,2,0,0,0]。

### 4.4.4 朴素贝叶斯分类

本文中将微博的情感分为三类，分别用数字代表某一类结果，其中1表示积极，2表示消极，3表示客观。经过之前的去标签、分词和向量化之后，样本均变成了numpy中的数组，下面将使用多项式朴素贝叶斯进行训练。其伪代码如下图4.4所示。

对人工标注中的类别进行计算

对于每篇训练的文本：

对于每个类别：

如果某个词语出现在文档中，增加其数值

然后再增加所有词语的计数值

对于每个类别：

除于总数目，得出条件概率

使用测试文档与词向量相乘，得出最大类别就是该文本的所属类别

图4.4 朴素贝叶斯

### 4.4.5 测试

之后使用从数据库中选取几个句子用来测试，结果如下（后面的标号表示分类结果，其中1表示积极，2表示消极，3表示客观）：

不惊扰别人的宁静，就是慈悲； 不伤害别人的自尊，就是善良。 ---- 3

与大哥太有缘分！竟是同天生日！！希望大哥的模特演绎之路能够越走越远越走越顺越走越好！！！ ---- 1

萌萌哒 ---- 1

现在的年轻人，连点小事都做不好 ---- 2

丁丁美人飞机临时取消，只能明天再见了[失望] ---- 2

默默看书 ---- 3

一天又一天，今夕是何年[失望] ----2

还好这只傻狗没有被学校的捕狗大队抓走[悲伤] ----2

相比游戏可以凭运气 ---- 3

生活中却全是坑，绕不开[哈哈] ---- 1

现实比不过一个网？[拜拜] ----2

最最亲爱的自己，最最亲爱的世界，晚安[兔子][月亮] ---- 3

有点想哭 ----2

发现只要过了星期三就变得好快[晕][晕] ----2

虽然下雨，健身不能停，瑜伽课加游泳八百米，代餐奶昔不裹腹啊，快睡吧，睡着了就不饿了！ http://t.cn/R2Wx9Wb ---- 3

### 4.4.6 计算准确率

本实验从训练集中随机选取100条用来测试，一共进行了十次实验，统计后将其正确率绘制成曲线如图4.4。

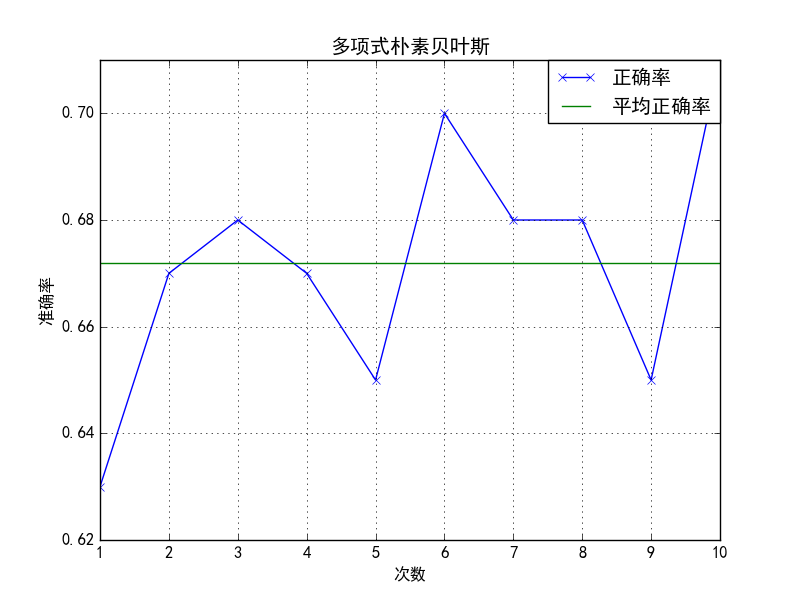


图4.5 准确率

## 4.5 本章小结

朴素贝叶斯，基于贝叶斯定理和特征条件独立假设的方法，目前在文本挖掘领域有了深入的研究。本文使用了标注之后的样本分为训练集和测试集，训练集朴素贝叶斯进行训练，然后对测试集进行预测，最后进行统计正确率。由实验结果看出，朴素贝叶斯的分类情况比较弱，下一步将使用AdaBoost对其进行分类的加强。

# 第五章 利用AdaBoost加强分类器

## 5.1 集成学习

所谓的集成学习，核心思想就是对数个单独的分类器进行加权，并将它们联合起来获得一个最强的分类器，其性能要比任何一个单独的分类器要强。该方法模拟了我们人类的特性，比如人类在进行一个重要的决策之前，都要搜索数个可能的判断，然后对这些判断进行加权，将它们联合起来形成一个最优的决策。

集成方法在很多领域都取得了广泛的应用。比如行星探测的轨迹增加、地震波的频率分析、web垃圾信息的过滤、生物唯一特征的识别等。在文本挖掘的领域，集成学习技术经常被用于情感分析、文本分类、倾向性分析等。

### 5.1.1 发展历史

集成方法的历史最早可追溯至1977年，由Tukeys Twicing提出了一个由两个线性回归模型构成的集成方法。两年之后，Dasarathy和Sheela提出用两个或多个分类器来对输入空间进行划分。20世纪90年代，Hansen和Salamon提出将多个类似配置的神经网络进行集成来提升单个神经网络的预测性能。目前的集成方法主要包括Bagging和Boosting算法。

### 5.1.2 Bagging方法

Bagging方法是由Breiman与1996年提出的，其基本思想是，将产生样本重复Bootstrap实例作为训练集，每一回运行Bgging都给学习算法提供替代的、随机从大小为m的初始训练集抽取出m个训练样本的集合。这种这种技术就叫做Bagging。一般情况下，Bagging要求分类算法不稳定，即样本集中的较小变动都能引起分类结果能够巨变，这些算法一般有决策树、神经网络等。

### 5.1.3 Boosting方法

Boosting的思想最早来源于1984年的Valiant。核心思想是，对样本赋予一个权重，分类精度高的弱分类器会有大的投票权重。Boosting算法主要包括两个分类：Boost-by-majority和Adaboost。

本文主要使用朴素贝叶斯作为AdaBoost的弱分类器，然后对分类器进行加强，最终选出一个最强分类器。

## 5.2 AdaBoost

Adaboost（Adaptive Boosting）也称自适应算法，最早由Freund和Schapine提出。该算法的主要思想是对那些很难分类的模式给予更多的关注。总关注度由一个权值来量化，然后赋予训练集中的样本。

### 5.2.1 概念

Adaboost是一种迭代的算法，会对同一个训练集使用不同的分类器训练，之后，再把这些分类器集合起来，构建一个最终的最强的分类器。其算法本身是通过改变一个权重D的分布来实现的，该权重D初始化一致，然后改变之后交给下一次分类器。使用Adaboost分类器能够过滤掉一些不必要的训练数据特征，然后放在关键的训练数据上面。AdaBoost分类流程图如图5.1。

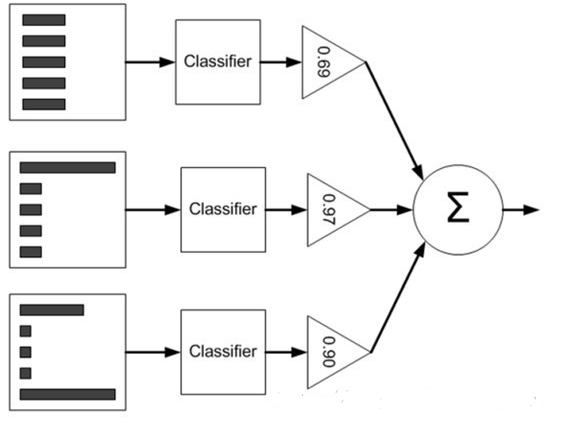


图5.1 AdaBoost流程

### 5.2.2 举例

下面我们举个例子来实现AdaBoost的实现过程：图中主要有“+”和“-”号，现使用决策树对其进行划分，主要将“+”和“-”号分割开来。

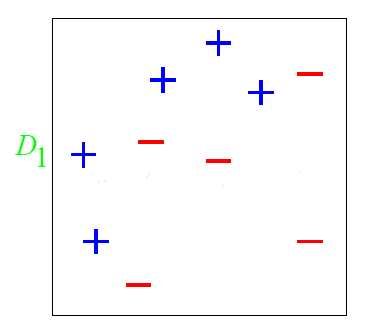


图5.2 实验集

初始分布采用均匀分布即。第一个分类器决策之后的边界如下图5.3。

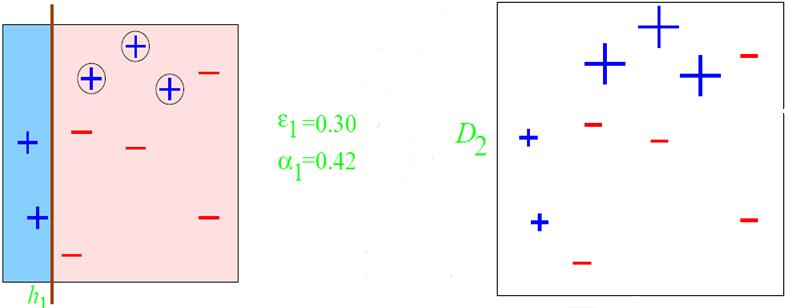


图5.3 第一次迭代

经过第一次分类之后，会得到一个新的样本权重，一个新的子分类器。但是，由图中可以看出，经过第一次划分之后，有三个“+”的被分错误。

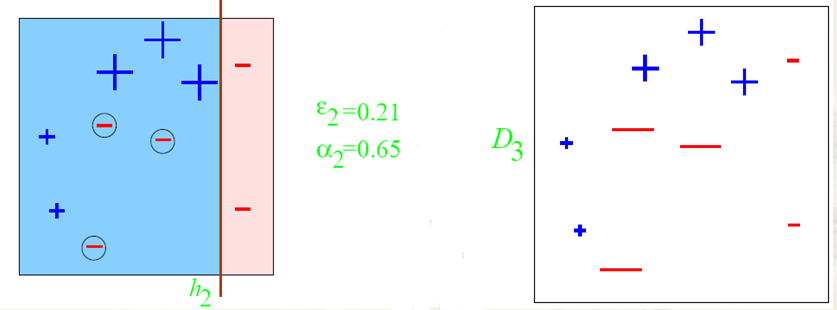


图5.4 第二次迭代

进行第二次迭代之后，再次得到一个新的样本权重D3，第二个子分类器h2。由图中5.4中可以看出，第二次分类仍是有三个“-”被分类错误。

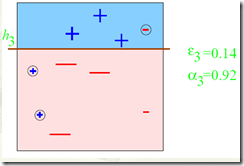


图5.5 分类器h2

第三次分类之后，其中有一个“-”和两个“+”被分类错误。再然后我们整合所有的分类器得出结果如下。

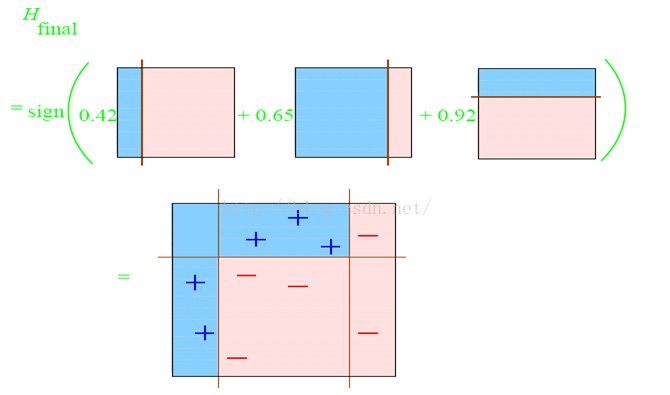


图5.6 将所有分类器结合

将三个分类器结合起来即可得出一个最强的分类器。

### 5.2.3 算法描述

令训练样本集D={(,), (,)…(,)}，，X用来表示训练样本集空间，是某一类别集。每次迭代的索引为t=1,2，…,T，AdaBoost算法在训练样本上维护一套权重分布W，其中，每个训练样本都对应一个权重，初始时，对所有i都有=1/N。伪代码如下：

1. input：N个训练样本{(,), (,)…(,)}，其中，类标签；

迭代次数T；

弱分类算法weaklearn。

1. 初始化：赋予每个样本相等的权重：=1/N。
2. for t=1 to T Do：
3. 在训练样本集D上，利用样本权重和weaklearn学习得到弱分类器：X;
4. 计算弱分类器的错误率：  
   ，I( （5.1）

If ：

退出循环

1. 计算：

（5.2）

1. 根据错误率更新样本的权重：

如果分类器结果正确即：if

（5.3）

如果分类器结果错误即：if

（5.4）

1. Output：最终分类器结果可以从中选出

（5.5）

整个算法的流程图如图5.7。



图5.7 AdaBoost流程图

## 5.3 AdaBoost提升朴素贝叶斯文本分类

将AdaBoost算法直接用于朴素贝叶斯算法的时候，在每次迭代过程中，训练样本如果被错误分类，权重将增加，否则将减少。

文本分类中，文本通常使用空间向量模型，即对于每个文本，都有一个特征向量=()，描述，其中，为该文本中提取出来的的权重。在使用朴素贝叶斯作为AdaBoost的基分类器时，分类过程中我们通常需要计算两个参数：

1. 文本属于每个类的先验概率；
2. 每个基于每个类的条件概率；

从上一章我们可以知道，朴素贝叶斯公式有：

（5.6）

其中，先验概率

（5.7）

而每个特征基于类的条件概率为：

（5.8）

其中，为特征出现的次数。如果文本属于类别时，，否则等于0。N和L分别代表的是训练文本数和类别数，V为训练文本集的特征集合。

在我们进行文本分类的时候，需要计算某个文本属于每个类别的后验概率，最后将其分到后验概率最大的一个类别。

（5.9）

防止乘积算出来的概率太低接近于0，故计算的时候取对数：

（5.10）

AdaBoost在迭代的时候，会为每个训练样本分配的权重为，然后将其引入参数，公式5.8随之会变为：

（5.11）

因此，随着AdaBoost 的每次迭代，样本权重每次都有更新，朴素贝叶斯的先验概率和后验概率都有变化，对朴素贝叶斯分类器的分类产生了扰动，增加了朴素贝叶斯分类器的相异性。

集成方法通过将多个分类器的分类结果组合起来，使最终分类分类器比最初的简单单分类器更好的结果。同时，多分类器能够解决单分类器的一些问题，比如过拟合问题。本文将使用朴素贝叶斯作为AdaBoost的基分类器，然后对微博进行情感分类。由实验可知，AdaBoost有效的提升了朴素贝叶斯的分类准确性。

## 5.4 多类问题

SVM、AdaBoost其本身是为了二元分类问题而设计的，但是在现实中的许多分类问题并不仅仅是二元分类，比如文本分类、人脸识别等，输入的数据都被划分为大于两个类。为了解决多元分类的问题，常用的两种方法如下：

1. 一对其他(OneVsRest)，即将多类问题划分成K个二类问题。将所有的数据属于某一类都看作主类，剩下的都看做负类。之后，再构建一个二元分类器。但是，如果一个样本被分到了正类，那么其他类都能够得到一票，显然，这种方法可能导致不同类的局。
2. 一对一(OneVsOne)，即构建K(K-1)/2个二类分类器，每个文本都只被只有这两类的分类器进行分类，剩下的全部忽略，以此循环，然后组合得到一个基于二元分类器的多分类方法。

针对多类问题，1996年，Yoav Freund在Experiments with a New Boosting Algorithm中提出了AdaBoost.M1和AdaBoost.M2两种算法。其中,AdaBoost.M1，也就是我们所说的Discrete AdaBoost。而AdaBoost.M2，则是M1的泛化形式。

### 5.4.1 AdaBoost.M1算法

伪代码如下：

已知：I表示弱诱导器；T表示迭代次数；S表示训练集。

需求：

（1）：令t=1

（2）：;

（3）：for t in T:

（4）： 利用I和分布

（5）： （5.12）

（6）： if then:

（7）： T=t-1；

（8）： 退出循环

（9）： end if；

（10）：

（11）： ；

（12）：将归一化为一个合适的分布；

（13）：结束，如果t>T.

最终的分类器是,当一个样本进入时,遍历所有类别,寻找使(样本属于某类y的情况下)最大者即是输出分类y。

AdaBoost.M1算法能够保证训练误差上届的指数下降，但是有个前提：基分类器的误差要小于0.5。与两类的AdaBoost方法相比，多分类的M1算法比较难找到准确率大于1/2的分类器。

### 5.4.2 AdaBoost.SAMME

对于多类的问题，AdaBoost.SAMME针对上诉两种问题，对AdaBoost的算法做了修改。算法通过改变权值的分配策略的计算方法，令：，SAMME与M1算法比较相似，其不同点就是中多加了个ln(k-1)项。但是，之前的研究发现，在处理很多分类问题上，SAMME算法的效果并没有M1的算法好。因为SAMME将分类器的准确率要求变为了1/k，不能保证训练样本分到本类的概率大于错分到其他各类的概率。

### 5.4.3 AdaBoost.SAMME.R

和SAMME算法一样，SAMME.R对分类器的性能要求都是只要比随机猜测好就能够接受。其算法代码如下：

已知：m个样本序列：{(,), (,)…(,)}其中为类别,且

步骤一：初始化权值。

步骤二：for t in T：

1. 按照权值，选择训练样本。
2. 对样本使用分类器进行分类：。
3. for k=1,2,…,K,执行下面两步：
4. 循环到各类中，分到各类样本的权值和：

for j=1,2,…,K

tjk=

1. 判断各分类的正确的样本权值之和是否大于其他类的样本的权值和。如果满足，则进行下一次循环，否则返回步骤二重新计算。
2. 计算分类器的伪错误率：

（5.13）

1. 然后调整alpha：

（5.14）

1. 重新计算权重向量：

（5.15）

1. 归一化

步骤三：最强分类器为：

（5.16）

在SAMME算法中，对分类器的只要求准确性大于1/k。在进行分类中，如果某个属于n类的样本，弱分类器将其分至某一不为n类的概率大于其分到n类的概率的情况会经常出现，经过无穷次迭代后，强分类器最终是会导致分类结果出现错误。SAMME.R会判断弱分类器的结果，判断正确分类的权值是否比分到其他任意一类的权值之和要大。如果满足则进进行下一次循环，否则进行重新计算。

## 5.5 实验

文本分类的类别可以有很多种分法，有的分为两类，如：积极和消极、主观和客观等，有的分为三类，如：积极、消极、客观，还有的分成四类，如：喜怒哀乐等。本文将对情感分类中的两类和多类进行探讨。

### 5.5.1 二分类

实验将训练集进行两类标注，分为积极和消极，经过文本预处理，然后采用朴素贝叶斯对训练集进行训练，其中使用AdaBoost对分类器进行加强。对于二分类，该算法的伪代码如下图5.8所示。由图中可以看出，经过50次迭代之后，AdaBoost有效的提升了朴素贝叶斯的分类准确性，由初始的27%变成了16%。

引入DS调整因子

对于每次迭代：

将DS与朴素贝叶斯的后验概率相乘

得出错误率

计算alpha

重新计算DS的权重向量

图5.8 二分类AdaBoost伪代码

实验结果：

根据错误率绘制图表如下图5.9：

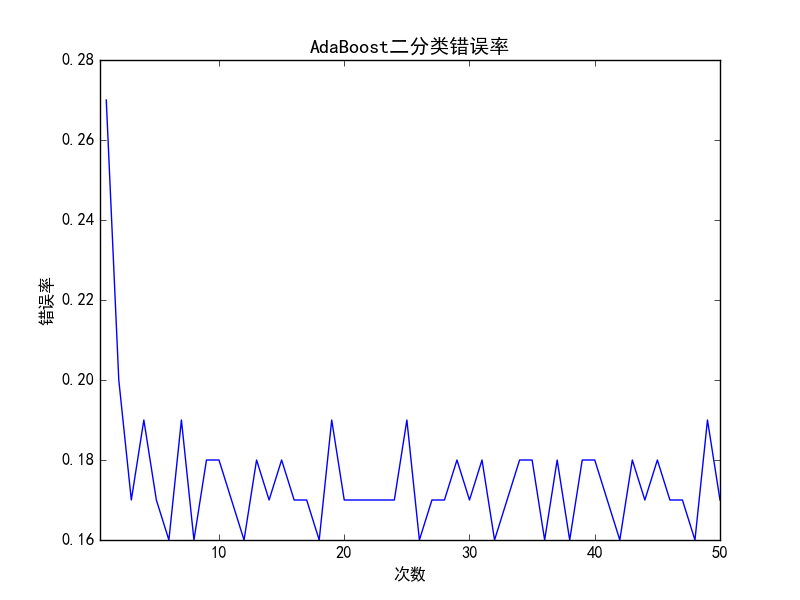


图5.9 错误率

同时，我们使用ROC曲线（受试者工作特征曲线）进行评测，主要多次不同二分类方式（分界值或决定阈），然后纵坐标表示真阳性率（灵敏度），横坐标表示假阳性率，理论上越靠近左上角表示分类越优秀。结果如下图5.10：

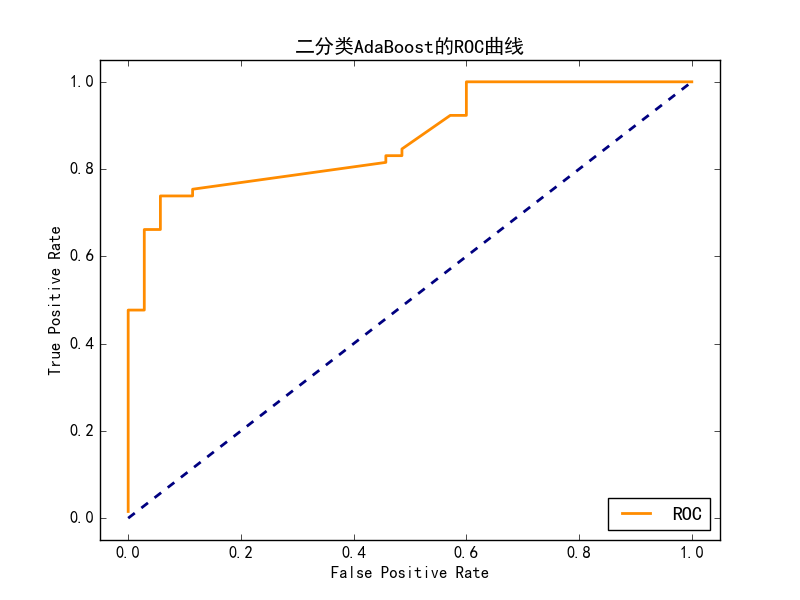


图5.10 ROC曲线

### 5.5.2 多分类

对于多分类的算法，AdaBoost的处理有多种方式，其中，以SAMME和SAMME.R效率较好。本文中将使用python sklearn库中的MultinomialNB和AdaBoost进行实验，sklearn优秀的实现了SAMME和SAMME.R算法，更与MultinomialNB能够完美无缝的结合。

1. SAMME算法

其调用方式如下：

ada\_discrete = AdaBoostClassifier(

base\_estimator=multi,

learning\_rate=learning\_rate,

n\_estimators=n\_estimators,

algorithm="SAMME")

本实验中，我们首先使用AdaBoost来对训练集进行自个训练（即对自己测试），之后，再用训练结果模型来对测试集进行预测。最后，我们将错误率的结果绘制成图如5.11所示。

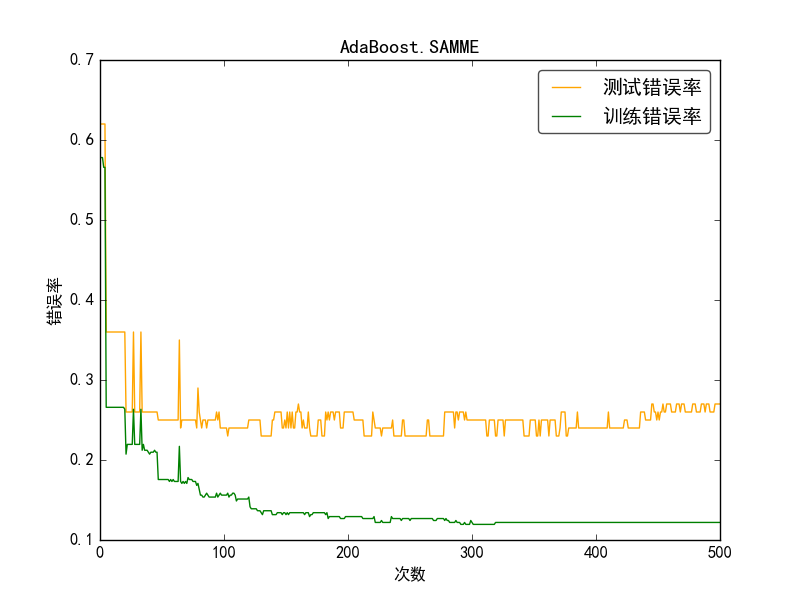


图5.11 AdaBoost.SAMME

由图中可以看出，SAMEE算法波动比较大，主要是由于如果分类错误，该分类算法会偏向于错分类别中概率较大的那一类，最终，分类器的正确率无法有效的提升。

（2）SAMME.R算法

SAMME.R算法的伪代码和SAMME中的类似，只是将AdaBoostClassifier中的algorithm换成"SAMME.R"。

ada\_real = AdaBoostClassifier(

base\_estimator=multi,

learning\_rate=learning\_rate,

n\_estimators=n\_estimators,

algorithm="SAMME.R")

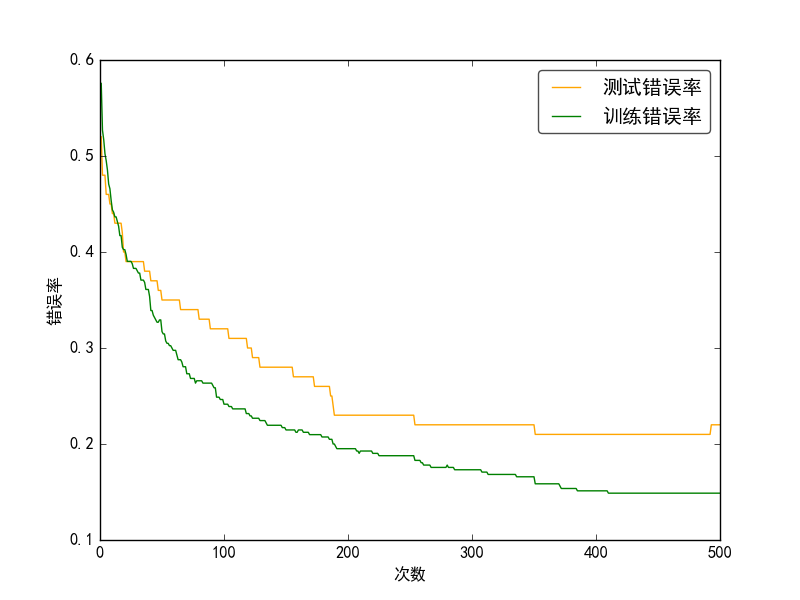


图5.12 AdaBoost.SAMME.R

由图中可以看出，SAMME.R算法比较平滑，波动较小。AdaBoost.SAMME.R是对AdaBoost的扩展和提升，其输出结果是一个实数值（也称置信度），将朴素贝叶斯作为其弱分类器之后，能够有效的提高实验的准确度。

## 5.6 本章小结

目前，AdaBoost算法的研究大部分都集中在分类问题，其中，人脸检测更是优于其他算法，其应用系列解决了多种分类的问题如：二分问题、多类单标签问题、多类多标签问题等。使用该算法仅仅需要增加新分类器，方式简单，同时能够让分类错误率的上届随着训练次数的增加而稳定下降，几乎不会产生过拟合的问题。本文将朴素贝叶斯作为AdaBoost的基分类器，处理多类问题，对微博的分类进行了加强，不仅防止了过拟合问题，还能降低了朴素贝叶斯的错误率。

# 第六章 总结与展望

微博已经成为了热点事件、体现大部分网民情感的新兴媒体，针对其的数据挖掘具有很广阔的前景和极其重要的研究价值。随着微博的高速发展，不规范、不规则的文本也越来越多，特别是网络流行词、使用国外文字等。本文结合了传统的文本分析方法并针对微博的特点，提供了清理无关性微博，之后再进行情感分析的方法。总体上，将微博分为了四类：无关类、积极、消极、客观。

## 6.1 本文主要内容总结

本文按照传统的文本分析方法，并根据微博的特点进行了情感分析。主要在以下方面展开了研究：

1. 对微博的降噪清理。从微博短文本的的特点，分析了采用Hash方法的URL并将其过滤，并将跟情感分析无关的用户名等进行过滤。进一步的，论文对微博中表达的情感基本单元也有深入的研究，通过对文本中的广义表情符号、重复单词等现象。
2. 对无关性的微博本文进行过滤。众所周知，微博已经为热点事件的传播平台，同时，也带来了大量无关的广告、重复字词（“分享图片”）等。本文利用了SVC对文本中无关性的微博进行了过滤。
3. 使用了朴素贝叶斯对微博进行情感分类。基于贝叶斯定理和特征独立的朴素贝叶斯方法，在文本分类中有极其明显的优点。
4. 使用AdaBoost算法对朴素贝叶斯进行加强。将朴素贝叶斯作为AdaBoost的弱分类器，然后对其进行加强。有效的提高了情感分类的准确性。

## 6.2 存在问题及未来研究展望

在微博的日益发展的情况下，还有很多方面有待开展进一步的研究。

1. 在情感分析的前提下，能够对某些微博中的评论来分析用户的情感倾向性，比如某些热点事件，分析大部分网民对热点事件的喜怒哀乐。同时，也可以根据该热点事件中牵涉到的时间、地点、人物等，对其深入的挖掘，甚至是做出预测性分析。这对保证社会稳定性有一定的作用。
2. 此外，也可更改情感分类的策略，以更精确的分析用户的语言现象，比如分析用户的程度副词如“非常”、“超级”等，结合文本中的标点符号和重复的词语，进行综合的整体建模。
3. 除了针对某些热点事件之外，还可获取个人所有的微博进行分析。从一个人的所有微博中可以获取其情感方向的估计，比如对某件事件的喜欢或者厌恶，对某些品牌的热衷与唾弃等。

总之，微博里已经涉及到了文本、图像处理、视频传播、社交网络等众多内容，是互联网时代中十分重要的媒体和平台。针对微博的深入研究，一方面可以对某个领域的基础和方法进行挖掘，另一方面，能够对社会产生巨大的价值，特别是维持社会稳定性方面。

# 参考文献

1. 《微博发布2016年财报：活跃用户增至3.13亿 全年营收首次突破10亿美元、净利润增长180%\_巨头\_投资界》.<http://news.pedaily.cn/201702/20170223409255.shtml>
2. Python爬虫和情感分析简介.<http://mt.sohu.com/20160504/n447633822.shtml>
3. 姚天昉 , 程希文 , 徐飞玉等.文本意见挖掘综述 [ J] .中文信息学报.2008, 22(3):71 -80.
4. 佘正炜.基于神经网络的文本倾向性分析系统的研究与实现[硕士学位论文].复旦大学.2012
5. 斯坦福大学讲怎样“情感分析”.<http://mp.weixin.qq.com/s?__biz=MjM5MTQzNzU2NA==&mid=202809946&idx=1&sn=95951e9f9b667f2ca77ac113565ad48d&>
6. 怎样利用Twitter/微博数据进行情感分析.<http://www.jianshu.com/p/dfcc59c4ce8c?ref=myread>.2017
7. 李毅.基于SVM的文本分类应用研究.[硕士学位论文].成都电子科技大学.2014
8. svm算法介绍 新浪博客.<http://blog.sina.com.cn/s/blog_a9303fd90101aj2a.html> 2012
9. 周志华.机器学习.北京：清华大学出版社.2016
10. 袁梅宇.数据挖掘与机器学习WEKA应用技术与实践（第二版）.清华大学出版社.2016
11. [美]Richard Lawson（李斌译）.用Python写网络爬虫.北京：人民邮电出版社.2016
12. [美] Ryan Mitchell（陶俊杰，陈小莉译）.Python 网络数据采集.北京：人民邮电出版社，2016
13. 张良均. Python数据分析与挖掘实战. 北京：机械工业出版社.2016
14. [美]Steven Bird，Ewan Klein，Edward Loper. Python自然语言处理.北京：人民邮电出版社.2014
15. 罗刚.自然语言处理原理与技术实现.北京：电子工业出版社.2016
16. [以色列]Lior Rokach（黄文龙 王晓丹 王毅 肖宇 译）模式分类的集成方法.北京：国防工业大学出版社.2015
17. [美]Peter Harrington（李锐 李鹏 曲亚东 王斌 译）.机器学习实战. 北京：人民邮电出版社.2013
18. 唐焕玲.基于半监督与集成学习的文本分类方法.北京.电子工业出版社.2013
19. 刘楠.面向微博的短文本的情感分析研究[博士学位论文].武汉.武汉大学.2013
20. 杨新武，马壮，袁顺.基于弱分类器调整的多分类AdaBoost算法.电子与信息学报.北京.北京工业大学.2016
21. JI ZHU,HUI ZOU,SAHARON ROSSET AND TREVOR HASTIE,Multi-class Adaboost, Statistics and its interface International Press,2009
22. 李航. 统计学习方法. 北京：清华大学出版社.2012
23. [以] 沙伊·沙莱夫-施瓦茨，[加] 沙伊·本-戴维 著；张文生 等 译. 深入理解机器学习：从原理到算法. 北京：机械工业出版社.2016

# 致 谢

经过了几个月对毕业设计的专研之后，本人基本完成了毕业设计。这次毕业设计，让我受益匪浅，不仅巩固了以往所学的专业知识，并且对于一些知识点有了更深层次的理解，特别是机器学习方面的知识，如SVM、朴素贝叶斯、AdaBoost等。

截至论文完稿之际，我要感谢王春东教授在多方面给予了我悉心的教导，比如论文选题、算法理解及论文撰写过程。在跟随王教授学习之时，更是对教授严格的教学态度、尽善尽美的工作作风、厚实的科研经验以及仔细负责的处事风范由衷地敬佩。

同时也要感谢一些同学在整个毕业设计期间给我的极大帮助，总是能够给我最好的解决方法，然后大家相互询问问题，相互帮助进步，完成整个论文设计。

最后就是谢谢我的家人，父母亲辛苦抚养我长大，在我任何有困难的时候都是对我鼎力相助，你们总是以你们无私的爱和一如既往的支持让我在求学的道路上继续前行,每每想起你们的这些付出都会让我充满了动力，所以我会在接下来的学习生涯里面继续发扬优秀的行为习惯。在学习和生活上做一个积极进取的人。

感谢所有曾经给予我帮助的老师、同学、亲人和朋友，我会继续努力。