分数:\_\_\_\_\_\_\_\_\_\_\_\_

2017－2018 学年 第一学期期末报告

《网络安全前沿技术》

实践讲堂研究报告

大数据舆情情感分析

学号:\_\_ZY1624134\_

姓名:\_\_\_是黎彬\_\_\_

学院:中法工程师学院

导师:\_\_荣文戈\_\_\_\_

2017年12月

目录

[摘要 3](#_Toc501283259)

[1. 引言 4](#_Toc501283260)

[1.1 情感分析概述 4](#_Toc501283261)

[1.1.1 背景 4](#_Toc501283262)

[1.1.2 问题描述 4](#_Toc501283263)

[1.2 课题内容 5](#_Toc501283264)

[2. 方法 5](#_Toc501283265)

[2.1 卡方检验特征选择 5](#_Toc501283266)

[2.2 支持向量机 5](#_Toc501283267)

[2.3 Pipeline 7](#_Toc501283268)

[3. 实验 7](#_Toc501283269)

[3.1 数据集 7](#_Toc501283270)

[3.2 数据处理与显示 7](#_Toc501283271)

[3.3 文本分类结果 9](#_Toc501283272)

[4. 总结 9](#_Toc501283273)

大数据舆情情感分析

是黎彬 ZY1624134 北京航空航天大学中法工程师学院 olivier.shi@buaa.edu.cn

摘要

随着互联网数据规模越来越大，开发一些模型去自动挖掘数据中的模式或者能够将数据与某些标签自动建立对应关系已经成为当前的重点和热点，尤其在情感分析领域。本课题在大数据舆情情感分析这个领域做了一定的研究与实践。本文在挖掘文本数据与人们情感之间的关联性时，首先观察数据的基本情况，考察与其相关统计信息，做一些针对性的处理。然后针对观察得到的信息，设计方法与模型，本课题是使用卡方特征选择方法去提取特征，即选择一些词为特征词，并把用特征表示过后的文本数据训练支持向量机分类器。最后在电影评论数据集上实践了我们的算法，并得到符合我的观察和想法的结果。

本文将在第一章概述情感分析的研究背景和意义以及本课题的研究内容；第二章将会介绍本课题涉及的方法，以及规划整个实验的流程图；第三章将会介绍本次课题在介绍方法上的实践，以及实验。最后总结以及未来可能继续进行的工作。

关键词： 卡方特征选择，支持向量机，文本分类，情感分析，大数据

1. 引言
   1. 情感分析概述
      1. 背景

在以信息技术为基础的革命推动下，以计算机、互联网、网络通信等信息技术为动力的信息化浪潮席卷全球。另外，随着社会技术的快速发展，来自博客、论坛、产品评价以及社会媒体的公共意见信息变得越来越多。网民也习惯于在做购买商品或购买股票的时候，从其他人中获得重要意见信息。例如，在购买一个商品的时候，越来越多的人首先从已有用户对商品的评价中做调查，从而获得重要的信息来决策。商业公司也紧密地关注自己的品牌、产品或服务，调查客户舆论，从而改善自己的产品质量。其中的情感分析或意见挖掘尤为重要，因为很多领域都需要去计算研究各自领域内用户或人民的意见和态度。在这样的背景下，做舆情分析是一个非常热门的也是急需的一项技术。随着互联网的发展和网民数量的增长，互联网上的数据量急剧膨胀尤其是文本数据的增长，先前的人工舆情分析手段远远跟不上如此快速的节奏。那么开发一种能够自动分析文本的情感倾向的技术就成了关键，也是近几年学术界和工业界的热点。

* + 1. 问题描述

情感分析任务是分析挖掘出人们日常的意见、态度和情感，通常情况下这个任务非常具有技术上的挑战性而且在实际上非常有用。情感分析通常包含如下四个任务：

1）实体抽取和聚类：从文本中抽取所有的实体并且将相同意思的实体表达出来。

2）方面抽取和聚类：从文本中抽取所有的实体的方面并且将相同的实体表达聚类。

3）意见持有者和时间抽取：从文本中抽取意见持有者以及对应的时间。

4）情感主客观分类：判断抽取出的实体方面是正极性还是负极性。

简而言之，情感分析就是对新闻资源、社会媒体评论和各大互联网社交平台上的网民意见等，带有情感色彩的主观或客观性文本进行分析然后处理进而归纳和推理的过程。本文主要研究情感信息的分类问题，即判断一个文本源是带有正面的评价，或者负面的反馈亦或中性的概述。

* 1. 课题内容

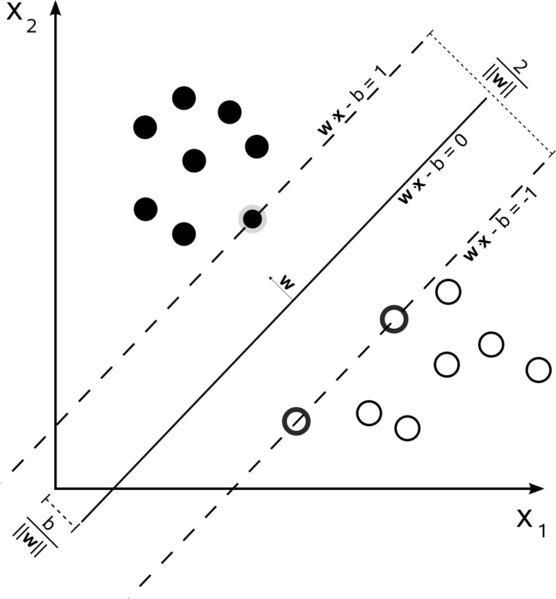
对于情感分析这个问题，可以尝试很多方法去做这个任务，包括最最传统的依靠构建情感字典的方法，到当前利用深度神经网络的模型去构建文本到情感的映射的建模方法。为了简单且有效的解决这个任务，本课题使用一个传统的机器学习方法——支持向量机（SVM），来当作文本分类器。另外，在具体实验的过程中，我们发现文本中有些词对于分类不起作用。不仅如此，当用支持向量机作分类器时，该分类是基于每个词的，那么如果词的数量过大，噪声过多，那么对于分类器来说，就越难从该数据中学习模式——从文本到类别的映射关系。对于这个问题，我们用卡方检验的方法去选择一些对于分类更加敏感且有效的词来作为分类的基本点，也就是文本表示特征（feature），再把这些挑选过的词作为特征训练支持向量机分类器模型。

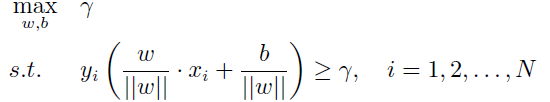
1. 方法
   1. 卡方检验特征选择

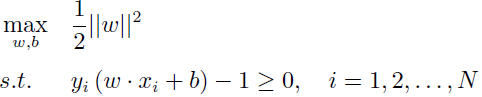
卡方检验最基本的思想就是通过观察实际值与理论值的偏差来确定理论的正确与否。理论值为E，实际值为X，那么偏差程度的就算公式为：

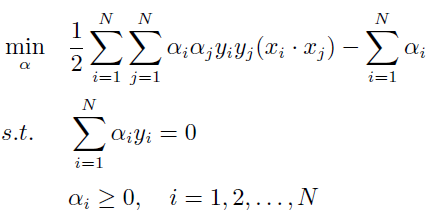
在文本分类的特征选择阶段，一般使用“词w与类别c不相关”来做原假设，计算出开方值越大，说明对原假设的偏离越大，即词与类别越相关。基于此，我们就可以选择那些开方值较大的词。

* 1. 支持向量机

支持向量机（SVM）是一个从感知器发展而来的分类模型。当训练数据集线性可分时，存在无穷多个分离超平面可将两类数据正确分开。感知器利用误分类最小的策略求得分离超平面，不过这时的解有无穷多个。而线性可分支持向量机利用间隔最大化求最优分离超平面，这时是唯一解。这里假设数据是线性可分的，在实际情况里，尤其是复杂的文本分类问题，这一假设并不成立，于是在SVM的基础上，又通过加松弛变量和核技巧等方法提升SVM的学习能力——使得SVM能够用某一维度的超平面尽可能将数据分开。这是SVM的整体思想与原理，图如下，

SVM的核心推导过程是一个优化问题转化的过程，即把非二次凸优化问题转变成二次凸优化的问题。首先从几何间隔最大化的问题，

转变成：

然后根据拉格朗日对偶性，该优化问题又能变为：

这就变成了一个二次凸优化问题，最后使用序列最小优化算法（SMO）来解这个优化问题。

* 1. Pipeline

结合上述介绍的卡方特征选择方法以及支持向量机分类器，我们给出以下流程。给定一个有标签的语料库（corpus），经过一些预处理（pre-process）后，把它切分成训练语料（train）和验证语料（dev），然后在训练语料上，用卡方特征选择（）方法选出卡方值较高的那些词作为特征，并用这些特征来表示训练和测试语料，以得到训练集和验证集。然后将将训练集去训练支持向量机分类器（SVM），用验证集去调试一些超参数（hyper-parameters），以得到最优的分类器

（classification）。最后可以用一些测试数据去测试一些该算法的性能。我们称该整体流程为基于卡方特征选择的SVM文本分类方法。

corpus

dev

train

feature

trainset

devset

SVM

classification

pre-process

1. 实验
   1. 数据集

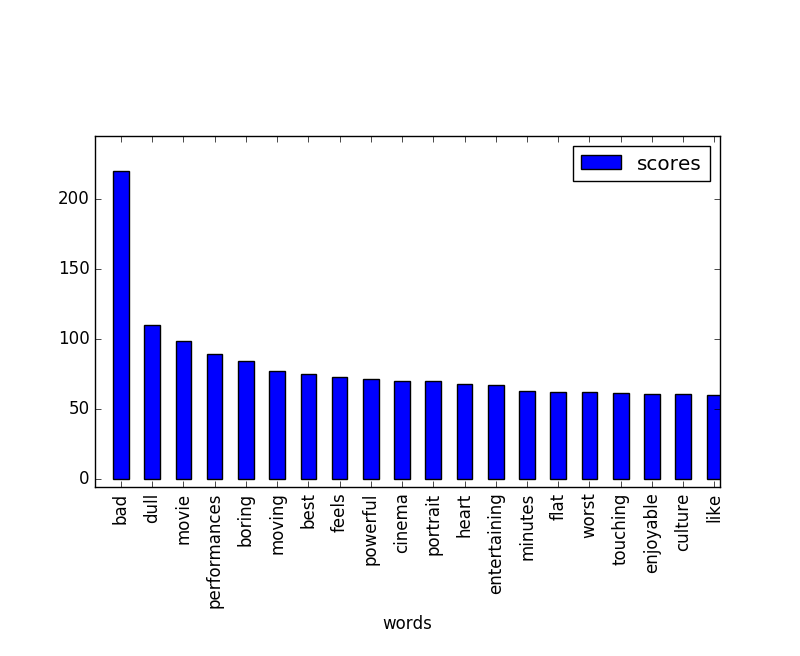
我们使用的数据集是电影评论数据集（Movie Review），该数据集有10624条评论，其中5312条表示情感为正，另外5312条表示情感为负。

* 1. 数据处理与显示

在拿到数据后，我们首先观察文本中词的分布特点。本人使用WordCloud工具将所有的正样本和负样本分别生成一个“词云”，如下图，左边为负，右边为正。



“词云”建图的特点是把出现频率高的词凸显出来，于是由图可以明显的看出一个特点，有很多词在正负样本中都出现的频率非常高，比如“movie, file, one, make, character…”，那么这些词对于文本分类可能就没那么重要。这也是本人要用一个特征选择方法的初衷。于是，我们用卡方特征选择方法计算了每个词的卡方值，从高到低对其进行排序并取出前20个词。我们发现，这里面有一些非常显著的带有情感色彩的词，如“bad, dull, boring, moving, best, enjoyable…”。这样的简单的可视化的分析过程也能够反映出方法的有效性。



* 1. 文本分类结果

上图表示，以带有高斯核的支持向量机为分类器，用卡方特征选择方法选择不同词数为特征数的分类准确率；其中，横轴表示选择不同的词数，本课题对比了（3000, 4000, 5000, 6000, 21282），21282是原始语料的所有词的数量；纵轴表示分类准确率。根据该表格显示，能够首先得出的结论是用卡方特征选择提升了分类器的准确率；此外，并不是说特征数越少就越好，这也是可以解释的，当特征数越少，那么就丢失了越多的词，那么该特征表示的延展性就很低，只能表示出含有这部分少数词的文本。考虑一种极端情况，就用卡方值最高的一个词“bad”作为唯一的特征词，那么如果一个文本中含有该词，那么该分类器可能做出比较准确的分类，但是如果文本中不包含这个词，这就使得分类器失效了。因此在实验过程中需要权衡选择词的数量。

1. 总结

本课题在大数据舆情情感分析这个领域做了一定的研究与实践。本文在挖掘文本数据与人们情感之间的关联性时，首先观察数据的基本情况，考察与其相关统计信息，做一些针对性的处理。然后针对观察得到的信息，设计方法与模型，本课题是使用卡方特征选择方法去提取特征，即选择一些词为特征词，并把用特征表示过后的文本数据训练支持向量机分类器。最后在电影评论数据集上实践了我们的算法，并得到符合我的观察和想法的结果。

在此次课题研究后，我还可以去调研用其他特征选择方法来选择词，如信息增益等方法，来对比各种特征选择方法的好坏；另外，我还可以尝试一些其他分类器，如逻辑斯蒂回归、决策树等分类器算法，并做一些对比实验。