浙 江 大 學

本 科 生 毕 业 论 文

题目 基于 SVM 的微博情感分析研究与实验

姓　　名 林炜华

学　　号 3110104065

指导教师 郑小林　副教授

专　　业 计算机科学与技术

学　　院 计算机学院

A Dissertation Submitted to Zhejiang University for the Degree of Bachelor of Engineering

TITLE: Research of Sentiment Analysis in Microblog Based on Support Vector Machine

Author: Weihua Lin

Supervisor: Asso.Prof. Xiaolin Zheng

Major: Computer Science and Technology

College: Computer Science and Technology

Submitted Date: 2015-6-03

浙江大学本科生毕业论文（设计）诚信承诺书

1.本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2.本人在毕业论文（设计）中引用他人的观点和参考资料均加以注释和说明。

3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

毕业论文（设计）作者签名：

_____年_____月_____日

摘要

微博作为最具代表性的在线社交网络，以其即时通信和社会化媒体优点吸引了数以亿计的用户群体，成为人们发表意见的重要载体。研究微博的情感有利于了解网络舆情，及时发现热门事件。以往的情感分析主要针对的是影评和产品评论，且以英文文本居多。

在对微博内容特点进行深入分析的基础上，我们提出了一种基于 SVM 的微博情感分析方法。我们采用手工标注语料集和 NLP&CC 2012 情感语料集，首先利用自然语言处理技术提取出影响微博情感极性的特征集，接着对不平衡特征集分布进行预处理，最后利用参数优化的 SVM 进行了训练和测试分析，并与朴素贝叶斯、随机森林和 K 近邻算法进行对比，发现参数优化的 SVM 在本实验数据集上的表现最好。之后，本文利用信息增益对各个情感特征的重要性进行了分析。

关键词 微博文本，情感分析，支持向量机，有监督学习，信息增益

Abstract

As a popular online social network platform, Microblog has attracted hundreds of millions of user groups with its characteristic of instant messaging and social media, thus becomes an important carrier for people to post views. Sentiment Analysis of microblog is conducive to understanding public opinions and discovering hotspot. Previous sentiment analysis researches mainly aim at movie reviews or product opinions, mostly in English texts.

In this article, we analyzed and extracted feature sets affecting sentiment of Sina microblog by natural language processing. Then we trained and tested preprocessed feature sets using SVM with optimized parameters on manual-labeled microblog datasets and NLP&CC2012 sentiment datasets. We also compared the performance of SVM with Naïve Bayes, Random Forests and KNN. Result turned out that SVM outperforms Random Forests and NB, while KNN proved to be the worst. Further, we analyzed the importance of each sentiment features using Information Gain.

Keywords Microblog, Sentiment Analysis, SVM, Supervised Learning, Information Gain

目录

第1章 绪论

1.1 情感分析背景

文本情感分析又称为意见挖掘，是指通过自然语言处理、文本分析、计算机语言学等技术对主观性文本进行分析、处理、归纳、推理的过程。

1.1.1 情感分析意义

随着互联网的飞速发展，在线用户不仅仅局限于信息的接收，而是更多地进行互联网信息的创造，包括对产品或服务的评论、对话题或事件的态度。快速准确地提取出文本中隐含的情感倾向，有利于用户在购买产品或服务时获取群众的评价从而帮助决策，也有利于政府部门及时追踪热点话题和突发事件的公众态度，了解舆情。网络中的情感信息不仅量大而且增长速度快，如果仅靠人工对文本进行整理识别将十分低效，不切实际。自然语言处理技术的研究成果对情感分析提供了强大的技术支持，采用计算机建模对大量文本进行情感分析成为了学术界的研究热点。

1.1.2 研究目的

微博作为最具代表性的在线社交网络，以其即时通信和社会化媒体优点吸引了数以亿计的用户群体，成为了人们发表意见，抒发情感的重要载体。对微博进行情感分析有利于了解网络舆情，掌握公众对话题或事件的整体态度。以往的情感分析研究主要集中在新闻评论、电影评论、产品评论等方面，且大多以英文文本为主。中文微博相对英文文本和普通中文文本来说包含了更多媒体信息，语法更加不规范，结构更加复杂，在情感特征的提取方面难度更大。因此本文目的是全面分析微博情感影响因素，制定情感特征提取方案，利用 SVM 等有监督学习方法对微博情感进行预测对比并分析各个特征对情感分类的重要性。

1.2 本文主要贡献

本文主要对以下两方面做出研究：

1) 分析微博文本情感极性的影响因素并提取相应特征

根据微博自身特点，结合自然语言处理技术（信息过滤、分词、词性标注等），分析微博的内容特征（词性、词的极性、词的极性强度、标点符号等）和媒体特征（@，表情等），基于情感词典，对手工标注集和公开的微博情感标注集进行特征提取，构建情感特征原始向量集。

2) 利用参数优化的 SVM 对情感特征集进行训练预测，并与随机森林、朴素贝叶斯和 K 近邻算法进行对比，最后分析情感特征的重要性。

对提取出的不平衡原始情感特征向量进行重新分布，利用参数优化的支持向量机（SVM）对特征集进行训练预测，得到精确率（precision）、召回率（recall）、F 值（F-score），之后比较了 SVM 和朴素贝叶斯（NB）、随机森林（RF）、K 近邻（KNN）在两个特征集上的效果，发现 SVM 的表现最好。之后，本文利用信息增益分析情感特征的重要性。

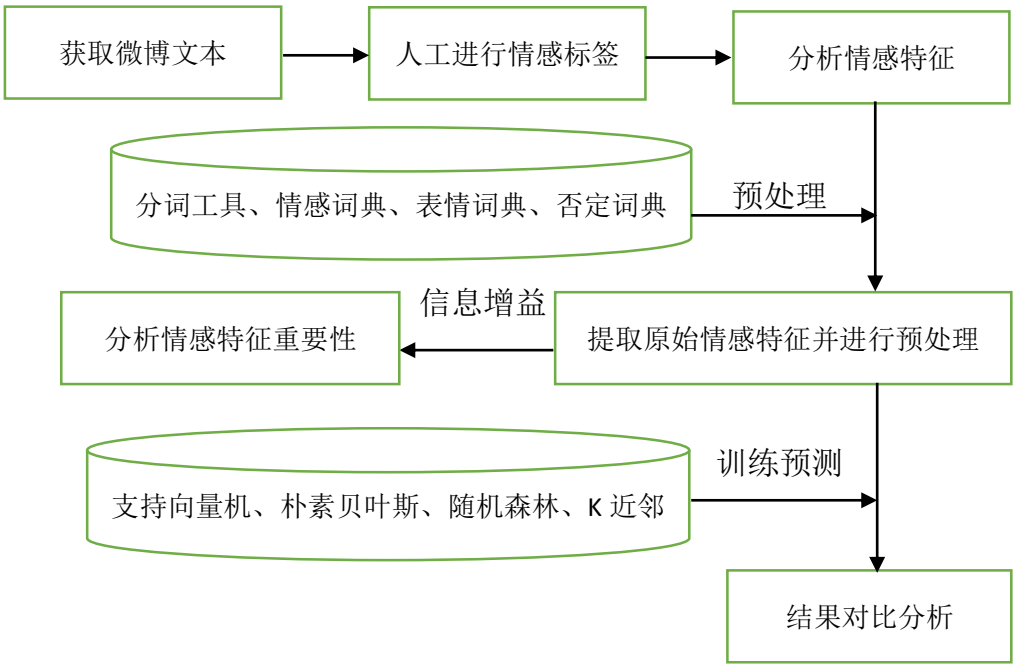

图 1-1 微博情感分析研究框架图

1.3 本文结构安排

本文共分五个章节，每章的研究内容和主要贡献如下：

第 1 章介绍了情感分析的研究背景、研究目的以及本文主要贡献。

第 2 章首先介绍了情感分析的相关背景，查阅相关文献并学习情感分析研究方法和模型特点，重点介绍了基于词典规则和有监督学习的情感分析方法，对中文微博的情感分析步骤进行了归纳整理，详述了本文用到的有监督分类算法的原理和优化。

第 3 章提出了本文的研究和实验方案。针对微博情感分析详述了基于情感词典的微博情感特征的提取方法（情感词典，情感表情、媒体特征预处理、分词、特征提取流程和算法）；详述了支持向量机的原理、流程和参数优化；详述了通过计算信息增益来衡量本文情感特征的重要程度。

第 4 章利用第三章中的特征提取方法对公开语料集和手工标注集提取相应情感特征向量，根据特征分布进行预处理并利用 SVM、NB、KNN、RF 四种有监督学习分类算法对特征集进行训练测试优化，对比不同情感分类模型的预测结果并对结果进行了分析，最后利用信息增益衡量了各个特征对情感分类的重要性。

第 5 章总结了本文的主要工作和存在问题，并对未来的工作进行展望。

第2章 情感分析文献综述

情感分析（Sentiment Analysis）指的是通过自然语言处理、文本分析和建立数学模型来识别和提取信息中的主观情绪。

情感分析具有巨大的实用价值，企业商家总是想要快速及时地了解公众和消费者对于他们产品或服务的评价和态度，从而进行产品服务的改进、合理营销和品牌宣传；潜在消费者在购买产品或服务时，同样想要了解以往用户对于该产品的评价考虑是否购买。当需要处理的信息量非常巨大时，人为进行调查将费时费力，此时利用情感分析能够及时准确地进行意见挖掘。

情感分析中的一个基本任务是对给定的文本进行极性分类——判断一个文档、一句话或者一个主题是正面、负面还是中立的。“超越极性“的情感分类则是试图判断出具体的情感状态，例如“伤心”、“愤怒”、“快乐”等。。

2.1 基于词典规则的方法

基于词典规则方法的思想是利用情感词典从文档中提取正负情感词来判断其极性。其中最容易实现的一种方法是在情感词典的基础上，提取文本中正向词语和负向词语的个数，根据正负向情感词词频来判断文本的极性，但是该方法对情感词典的质量要求极高。情感词典的构建一般是利用已标注情感词作为参考，计算新词和已标注词的语义相似度来划分词性。

Riloff 等人[1]提出了一种通过构建情感模板（bootstrapping）对文本极性进行判断的模板分类方法，运用大量未标注的语料识别主观句。由于模板制定局限性较大，所以这种方法不能够保证全面性和有效性。Hu 等人[2]以 WordNet 中的同义词和反义词为基础得到词语的情感倾向，从产品的用户评论中提取出正面和负面的情感句，得到了较高的精确率和召回率。但是他们只使用了形容词作为句子的语义导向，没有考虑动词和名词，另外他们没有对情感句的情感强度进行研究。

Turney 等人[3]提出了一种基于平均语义导向（Semantic Orientation）的非监督算法，将 epinion 上关于手机、银行、电影以及旅游目的地相关的评论分类为“推荐”或者“不推荐”。一个评论由多个包含形容词或动词的短语构成，短语的语义导向是通过计算它与词“excellent”的点互信息和它与词“poor”的点互信息

之差得出的。如果一条评论的平均语义导向为正，则该评论视为“推荐”，否则视为“不推荐”。 Saif 等人[4]提出了一个基于词典的使用词语上下文进行表达的 SentiCircle 方法，能够从词语的上下文一致性中捕捉到隐含的语义信息，并由此更新情感倾向。他们在三个 Twitter 数据集上使用了三个不用的情感词典评估了 SentiCircle 方法，发现 SentiCircle 方法比其他基于词典的方法精确度更高。

2.2 机器学习的方法

机器学习方法的主要思想是先使用情感词、情感短语等作为分类特征，然后利用朴素贝叶斯（Naïve Bayes）、最大熵（Maximum Entropy）、支持向量机（Support Vector Machine）、决策树（Decision Tree）、K 最近邻（K Nearest Neighbor）等分类模型，在人工标注过情感极性的样本集合上训练，最后利用训练好的分类模型对测试集的极性进行预测。

Pang 等人[5]对电影评论文本进行否定词、一元词、二元词、词性标注等特征提取，然后利用提取出的特征，分别使用 SVM、Max Entropy 和 Naïve Bayes 三种不同的分类算法来进行情感极性的分类。实验结果表明使用一元模板（unigram）的词袋作为分类特征在朴素贝叶斯和 SVM 模型上效果较好。Whitelaw 等人[6]基于评价理论将一个评价组（appraisal group）表示为几个独立的语义分类中的属性值集合，利用半监督方法建立了评价性形容词的词汇表以及相关修饰语，利用向量空间模型和 SVM 对电影评论进行正面和负面的分类，得到了极高的准确率。SVM 在情感分析中被广泛地使用并取得了不错的效果，而人工神经网络（Artificial Neural Networks）在情感分析研究中则极少出现。为此，Moraes 等人[7]在文档级别的情感分析中将 SVM、ANN、NB 进行了实验的对比，在传统的词袋模型中采用了带有监督的特征选择和权重计算方法的标准评估文本。实验结果表明，除了一些不平衡的数据文本，ANN 算法相对 SVM 来说效果不相上下，尤其是在电影评价的标准测试集中，ANN 精确度要高于 SVM，而 SVM 对于噪声项的适应要强于 ANN。同时，实验还证明了两个模型在情感分类应用中的潜在限制，就像预期的那样，ANN 的训练时间要远远大于 SVM；而相反地，SVM 由于支持向量数量庞大，在预测时间上要大于 ANN。

情感分析中的机器学习方法准确性较高，但是对训练语料依赖性大，训练周期相对比较长。

2.3 中文微博情感分析

微博作为最具代表性的在线社交网络，以其即时通信和社会化媒体优点吸引了数以亿计的用户群体，成为了人们发表意见的重要载体。对微博进行情感分析有利于了解网络舆情，掌握公众对话题或事件的整体态度。

相对英文来说，中文文本更加复杂，分词过程更加困难。而相对传统文本来说，中文微博包含的文本信息更加丰富，有着新的多特征符号加入，需要重新考虑。总的来说，中文微博的情感分析过程可以分为预处理、情感信息提取和情感分类几个阶段。

文本预处理包括分词、停用词去除、词性标注等步骤，常用的分词系统有中国科学院计算技术研究所所研制的基于多层隐马尔科夫模型的汉语词法分析系统 ICTCLAS[8]（Institute of Computing Technology, Chinese Lexical Analysis System）。ICTCLAS 系统分词速度快，分词精度高，不仅能对文本进行分词，还能对分词结果进行词性标注。哈工大社会计算与信息检索研究中心研制的 LTP[9]（Language Technology Platform）开源语言技术平台包含了分词、词性标注、句法分析等基于 XML 的中文原因处理模块。除了分词等工作，还需要针对微博特点对链接、主题标签、@等媒体信息进行过滤。

情感信息提取方面，朱嫣岚等人[10] 基于 HowNet 中的少量基准词，提出词汇语义相似度和语义相关场的情感词极性计算方法，在常用词集的测试中达到了 80%以上的正确率。徐琳宏等人[11]对情感词的极性和强度进行了整理，在手工情感分类和强度分类基础上，通过计算待定词汇和手工标注词汇的点互信息决定待定词汇的情感强度，构建情感词汇本体库。

情感分类方面，刘志明等人[12]使用支持向量机（Support Vector Machine）、朴素贝叶斯（Naïve Bayes）、最大熵（Maximum Entropy）三种机器学习方法，信息增益（IG）、CHI 统计、文档频率（DF）三种特征选取方法以及布尔型特征权重（Presence）、词频型特征权重（TF）、TF-IDF（Term Frequency-Inverse DocumentFrequency）三种特征项权重方法对微博进行了情感分类，发现采用 SVM、IG、以及 TF-IDF 三者结合的方式对微博的情感分类准确率最高。

2.4 有监督学习算法详述

在统计学中，分类问题的目标是根据历史观测数据的类别，识别出新的观测

数据属于哪一类。分类通常可以分为有监督学习以及无监督学习，有监督学习通过训练已标记样本集产生一个最优分类模型，应用于新观测数据的类别预测；无监督学习直接对未知标签样本集进行建模，最典型的例子是聚类。常用的有监督学习方法有支持向量机、朴素贝叶斯、K 最近邻、随机森林等，下文将会对以上有监督学习方法进行介绍。

2.4.1 支持向量机

SVM(Support Vector Machine)[14]是 90 年代中期发展起来的基于统计学习的二分类模型，其基本模型定义为特征空间上的最大间隔线性分类器，通过距离最大化来降低结构风险。

2.4.1.1 线性分类器

对于一个简单的二分类问题，可以用 n 维向量 x 来表示数据点，线性分类器目标是在 n 维数据空间中找到一个超平面对数据进行准确划分，公式表示为：

$$w^T\mathrm{x} + \mathrm{b} = 0 \tag{2.1}$$

2.4.1.2 逻辑回归

逻辑回归(Logistic Regression)[15]使用逻辑函数（也称 Sigmoid 函数）将特征的线性组合作为自变量映射到（0,1）上，目的是从特征学习出 0/1 分类模型，映射后的值就被认为是 y = 1 的概率。逻辑函数（图 2-1）表示为：

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

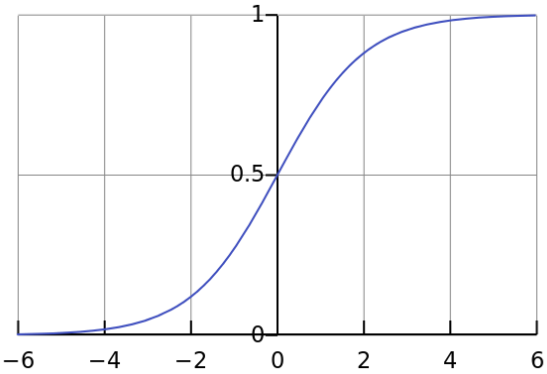

图 2-1 逻辑函数

对于二分类问题，假设函数就是特征属于 y = 1 的概率，可以得到：

$$\mathrm{P}(\mathrm{y} = 1|\mathrm{x};\theta) = h_\theta(x) \tag{2.2}$$

$$\mathrm{P}(\mathrm{y} = 0|\mathrm{x};\theta) = 1 - h_\theta(x) \tag{2.3}$$

当判别一个新来的特征属于哪一类时，只需求$h_\theta(x)$，大于 0.5 即为 1 类，否则为 0 类。

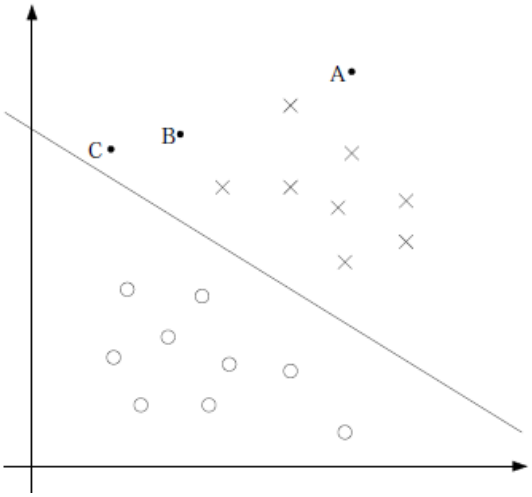

图 2-2 线性分类超平面

上图中假设通过逻辑回归学习出的分类面为$\theta^T x = 0$，从图中我们可以确定 A 点为 X 类，但是对于 C 点却不太确定。从分类来说，我们应该更加关注靠近中间分割线的点，让他们尽可能远离中间线，而不是在所有点上达到最优。

2.4.1.3 最优间隔分类器

给定训练样本(x^i, y^i)，x 为特征，y 为标签（-1 和 1），i 表示第 i 个样本。几何间隔定义为：

$$\gamma^{(i)} = y^{(i)}((\frac{w}{||w||})^T x^{(i)} + \frac{b}{||w||}) \tag{2.4}$$

全局样本上的几何间隔定义为：

$$\gamma = min_{i=1,\dots,m}\gamma^{(i)} \tag{2.5}$$

SVM 的目标是寻找一个超平面，使得离超平面最近的点具有最大间距。形式化表示为：

$$\begin{aligned} & max_{\gamma,\mathrm{w},\mathrm{b}} \ \gamma \\ \text{s.t.} \ & y^{(i)}(\mathrm{w}^T x^{(i)} + b) \geq \gamma, \ \mathrm{i} = 1, \dots, \mathrm{m} \\ & ||\mathrm{w}|| = 1 \end{aligned} \tag{2.6}$$

在此约束$||\mathrm{w}|| = 1$，使得$\mathrm{w}^T x^{(i)} + b$为几何间隔。

函数间隔定义为：

$$\hat{\gamma}^{(i)} = y^{(i)}(\mathrm{w}^T x^{(i)} + b) \tag{2.7}$$

全局样本上的几何间隔定义为：

$$\hat{\gamma} = min_{i=1,\dots,m}\gamma^{(i)} \tag{2.8}$$

由于几何间隔$\gamma = \frac{\hat{\gamma}}{||w||}$，因此可以把公式(2.6)转化为：

$$max_{\gamma,w,b} \ \frac{\hat{\gamma}}{||w||}$$
$$s.t. \ \ y^{(i)}\left(w^T x^{(i)} + b\right) \geq \hat{\gamma}, \ i = 1, \ldots, m \qquad (2.9)$$

由于同时扩大 w 和 b 对结果没有影响，因此我们可以将全局函数间隔$\hat{\gamma}$定义为 1，即将离超平面最近的点的距离定义为$\frac{1}{||w||}$，由于求$\frac{1}{||w||}$的最大值相当于求$\frac{1}{2}||w||^2$的最小值，因此将公式(2.9)改为：

$$min_{\gamma,w,b} \ \frac{1}{2}||w||^2$$
$$s.t. \ \ y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \ i = 1, \ldots, m \qquad (2.10)$$

于是，如图 2-3 所示，使得$||w||^2$最小的分类面就称为最优分类面，B1，B2 上的训练样本点就称为支持向量。

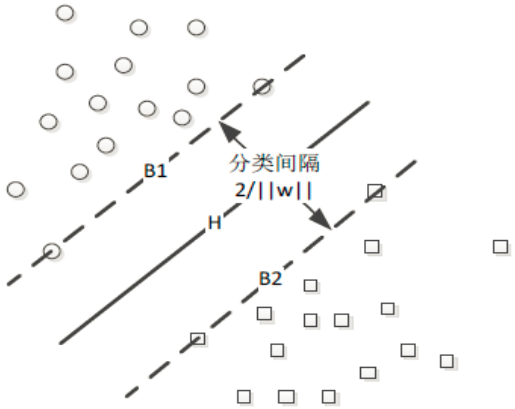

图 2-3 最优间隔分类

2.4.2 朴素贝叶斯分类

朴素贝叶斯(Naïve Bayes)[16]是一种基于统计学习的简单分类算法，广泛应用于文本分类等领域。它的基本思想是利用特征项分别计算文档类别的联合概率，然后取所有计算出的条件概率中最大的那个类别作为文档的类别。朴素贝叶斯分类的过程如下：

1) 设 x=$\{a_1, a_2, \ldots a_m\}$为待分类项，a_i为 x 的特征属性
2) 有类别集合 C=$\{y_1, y_2, \ldots y_n\}$
3) 对于每个类别分别计算 P(y_i|x)
4) 如果 P(y_k|x) = max{P(y_1|x) , P(y_2|x) , ... P(y_n|x) }，则x ∈ y_k

对于步骤 3 中的条件概率：

1) 利用训练数据得到每个属性的条件概率：

$$\mathrm{P}(a_1|y_1), \mathrm{P}(a_2|y_1), \dots \mathrm{P}(a_m|y_1); \mathrm{P}(a_1|y_2), \mathrm{P}(a_2|y_2), \dots \mathrm{P}(a_m|y_2); \dots ; \mathrm{P}(a_1|y_n), \mathrm{P}(a_2|y_n), \dots \mathrm{P}(a_m|y_n) \quad (2.11)$$

2) 根据贝叶斯公式：

$$\mathrm{P}(y_i|\mathrm{x}) = \frac{\mathrm{P}(\mathrm{x}|y_i)P(y_i)}{P(x)} \quad (2.12)$$

假设各特征属性是条件独立的：

$$\mathrm{P}(\mathrm{x}|y_i)P(y_i) = \mathrm{P}(a_1|y_i)\mathrm{P}(a_2|y_i) \dots \mathrm{P}(a_m|y_i)\mathrm{P}(y_i) = \mathrm{P}(y_i)\prod_{j=1}^{m}\mathrm{P}(a_j|y_i) \quad (2.13)$$

贝叶斯分类理论上具有最低的分类错误率，但是朴素贝叶斯假设"各个特征属性彼此独立"，在现实中往往不太可能成立，会导致模型与实际的误差。

朴素贝叶斯分类具体算法如下：

```
Input：训练集 C={C1,C2,…,Ck},F={F1,F2,…,Fn},train_label{L1,L2,…Ln}
       测试集 testF={testF1,testF2,…,testFn},test_label{TL1,TL2,…,TLn}
       其中 Fi={Fi,1,Fi,2,…,Fi,m}, testFi={TFi,1,TFi,2,…,TFi,m}
Output：测试集预测结果 predict_label
//训练部分
For each class Ck:
    Compute P(Ck);      //计算决策属性各分量概率
    For each training feature Fi:
        For each conditional attribute Fi,j:
            Compute P(Fi,j|Ck);      //计算后验概率
        End
    End
End
//预测部分
For each testing feature testFi:
    For each class Ck:
        //计算样本对每一类的后验概率之积和该类概率的乘积
        Compute P_Xk = P(TFi,1|Ck) P(TFi,2|Ck)…P(TFi,m|Ck)P(Ck)
    End
    predict_label(i) = class(max(P_Xk));   //将 testFi 归类为 P_Xk 最大的类
End
```

朴素贝叶斯要求计算特征划分的条件概率，当特征值为离散值时，只需统计特征的各个值的频率即可。当特征值为连续值时，一方面可以对特征进行分段处理；另一方面可以假设特征值服从高斯分布，计算训练样本中特征项划分的均值和标准差，代入高斯函数估计概率。

在实际情况中，有时候某一类中某个特征项的某个值并没有出现过，这样将会使得测试数据在该类的条件概率为 0，为了解决特征项划分时出现 0 概率的问题，我们引入 Laplace Estimation，对出现 0 频率的划分的计数加 1，如果训练样本数量充分大时，并不会对结果产生影响。

2.4.3 K 近邻

K 近邻(K-Nearest Neighbors)[17]分类是基于特征空间距离的分类方法。它的基本思路非常简单：对于未标签数据，计算它和每个已标签训练数据的空间距离，考虑离它最近的 K 个数据的分类情况，将频数最多的那个类视为数据的分类。如图 2-4 所示，当 K 为 3 时，离圆形最近的三个数据有两个为三角形，因此圆形归为三角形类；当 K 为 5 时，离圆形最近的五个数据有三个为正方形，因此圆形归为正方形类。

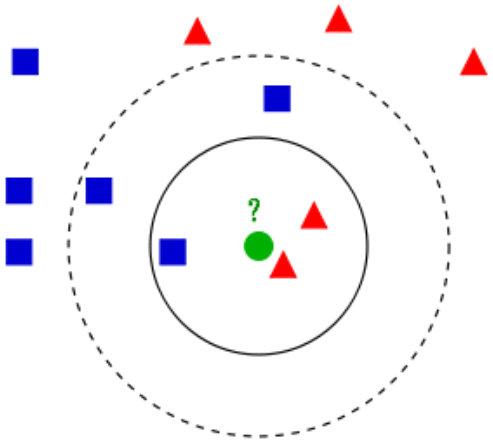

图 2-4 K 近邻分类示意图

KNN 分类主要靠邻居来决定所属类别，对于重叠较多的样本数据 KNN 能够发挥更好的效果。KNN 分类有两个缺陷：KNN 算法对数据的局部结构非常敏感，当样本分布不平衡时，出现频率大的样本将会主导测试点的测试结果；此外，KNN 属于计算密集型，每次分类都要计算数据与所有已标签数据的距离，因此计算量较大。对于 K 近邻算法，需要尝试不同的 K 值选取准确率最大的 K 作为模型的参数。

2.4.4 随机森林

随机森林(Random Forests)[18]由一定数量相互独立的决策树组成，它的基本思路是：当输入一个新的测试样本时，每一棵决策树分别对该样本进行分类，最后选择众数最大的类别作为测试样本的分类结果。

2.4.4.1 决策树

决策树是一个基于规则的批处理预测模型，它表示样本特征与样本类别的映射关系，通过树的分支对数据依靠特征进行分类。决策数中每个节点代表一个特征，节点的分叉代表特征值的不同划分，叶节点代表分类结果。

决策树的构造过程中要考虑节点分裂属性的选择以及树剪枝。分裂属性的选择即在 n 个特征属性中，按照特征的重要性对特征进行优先筛选，分裂属性的选取方法主要有 ID3、C4.5、CART 三种。树剪枝目的是解决过拟合问题，在构建树叉时，数据的噪声反映了训练数据中的异常，树剪枝通常分为先剪枝和后剪枝。

2.4.4.2 随机森林

随机森林由多棵 CART（Classification and Regression Tree）构成。在建立每一棵决策树过程中，训练集是从总的训练集中有放回采样的，因此某些训练样本可能重复出现在相同树的训练集中，也可能从未出现。这样每棵树的训练样本都不会是总的样本，不容易出现过拟合问题。在训练每棵树的节点时，通常以 sqrt(M) 左右的比例随机无放回抽取 m 个特征进行这棵树的训练（M 为特征总数），这样每棵树就类似一个精通某个窄领域的专家，分别从不同的角度（m 个子特征）去看待测试数据。预测时每一棵树都对测试数据进行决策，最后将所有决策树的投票结果综合选取类别最多的那一类。

对于随机森林，决策树数量的多少对分类的效果有重要的影响，对于大量数据来说随着 CART 的增大训练时间也更加长，到达一定数量之后准确率就不会有明显的提升了，因此需要尝试不同的决策树数量对比准确率的变化，最终确定合适的数量

2.5 本章小结

本章主要介绍了情感分析的相关背景，查阅相关文献并学习情感分析的研究方法和模型特点。重点介绍了基于词典规则和有监督学习的情感分析方法，对

中文微博的情感分析步骤进行了归纳整理。

在有监督学习算法详述部分阐述了本文所用到的分类算法的原理和相关的数学理论。重点介绍了支持向量机、朴素贝叶斯、K 近邻以及随机森林四类有监督学习分类方法以及参数优化。

第3章　基于 SVM 的微博情感分析方法

3.1 概述

微博情感分析：获取新浪微博相关主题的文本数据进行人工标注情感和预处理，然后参照情感词典、表情词典抽取微博的内容特征和媒体特征，根据特征分布对不平衡特征集进行分段处理，利用参数优化的 SVM 对特征集进行训练预测，得到精确率、召回率和 F 值，与 NB、RF、KNN 三种有监督学习算法进行对比，分析结果。最后利用信息增益计算各个情感特征的重要性。

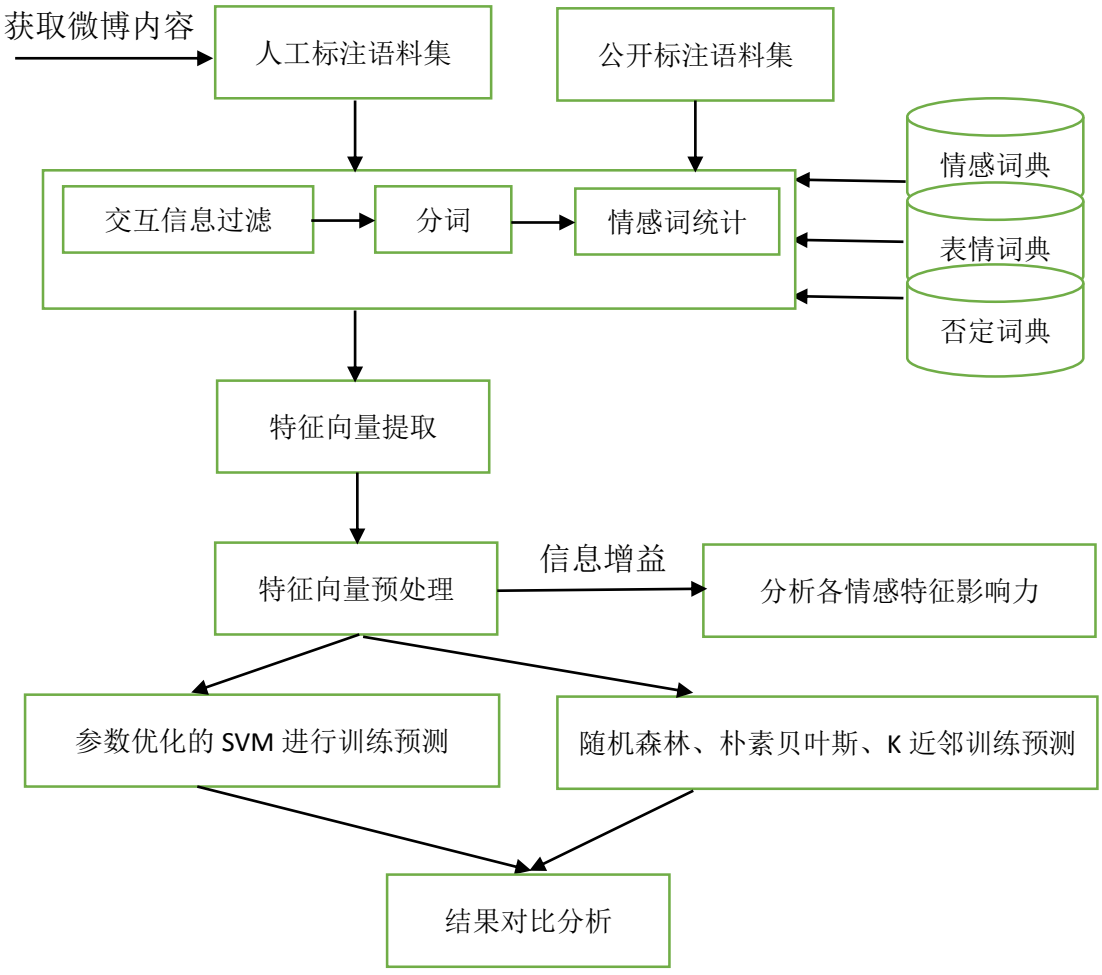

图 3-1 总体研究框架图

3.2 基于情感词典的微博情感特征提取

情感特征提取是对微博进行多媒体信息预处理和分词，利用情感词典、表情词典等情感库提取出微博中情感特征的过程。

3.2.1 情感词典和表情库

情感词典是情感分析的基础，常用的中文情感词典主要有知网整理的情感分析词语集、台湾大学自然语言处理实验室（NTUSD）整理的中文情感极性词典以及大连理工大学信息检索研究室整理的中文情感本体库。

中文情感词库包含情感词 27466 个，将情感分为 7 大类 21 小类，并且对情感词的强度分为 1,3,5,7,9 五档，1 为强度最小，9 为强度最大。本文就是利用中文情感词库作为情感词典，情感词分类如表 3-1 所示：

表 3-1 中文情感词库情感分类

编号	大类	情感	例词
1	乐	快乐	喜悦、欢喜
2		安心	踏实、宽心
3	好	尊敬	恭敬、敬爱
4		赞扬	英俊、优秀
5		相信	信任、信赖
6		喜爱	仰慕、宝贝
7		祝愿	渴望、保佑
8	怒	愤怒	气愤、恼火
9	哀	悲伤	忧伤、悲苦
10		失望	憾事、绝望
11		疚	内疚、忏悔
12		思	思念、相思
13	惧	慌	慌张、心慌
14		恐惧	胆怯、害怕
15		羞	害羞、害臊
16	恶	烦闷	憋闷、烦躁
17		憎恶	反感、可耻
18		贬责	呆板、虚荣
19		妒忌	眼红、吃醋
20		怀疑	多心、生疑
21	惊	惊奇	奇怪、奇迹

微博的表情往往能够直观地反映出发布者的心情，因此对微博表情极性进行

整理十分重要。本文将微博表情分为正面，中性和负面三类，其中正面表情 103 个，负面表情 74 个，中性表情 91 个，具体如表 3-2 所示：

表 3-2 微博表情极性分类表

	个数	举例
正面表情	103	［哈哈］、［偷笑］、［爱你］、［心］、［嘻嘻］、［鼓掌］、［给力］、［可爱］、［威武］…
负面表情	74	［泪］、［怒］、［衰］、［生病］、［失望］、［委屈］、［鄙视］…
中性表情	91	［吃惊］、［兔子］、［围观］、［思考］、［浮云］、［话筒］、［疑问］、［风］…

3.2.2 预处理和分词

微博文本相对传统文本有着较大区别，主要体现在：文本篇幅较短；语法较为不规范；包含链接、表情、标签、交互信息等媒体特征。

在进行分词之前，必须先利用正则表达式制定过滤规则对微博内容进行预处理，主要有以下几个方面：

1) 媒体特征过滤。链接 URL、标签 Hashtag、呼叫@等媒体特征过滤。
2) 标点符号统计。微博中的连续标点符号、问号以及感叹号都可能表达了微博的情感极性，需要进行记录。
3) 表情符号统计。微博文本中的表情以“[表情]”的形式出现，表情对情感的表达有着重要的反映，本文根据整理的微博表情极性分类库对微博中正面、负面、中性三类表情进行统计。

中文语言相对英文来说更加复杂，分词难度也更大。目前比较成熟的中文分词技术有基于隐马尔科夫模型分词算法、基于 n-最短路径分词算法、最大匹配算法等。

本文的微博分词采用了中国科学院计算机研究所汉语词法分析系统 ICTCLAS(Institute of Computer Technology, Chinese Lexical Analysis System)作为辅助工具。该分词系统分词速度快，分词精度高，不仅能对文本进行分词，还能对分词结果进行词性标注。形容词、名词以及动词的词性标记集如下表所示：

表 3-3 ICTCLAS 形容词、名词、动词词性标记集

词性（总）	符号	词性
名词	n	名词
	nr	人名
	nr1	汉语姓氏
	nr2	汉语名字
	nrj	日语人名
	nrf	音译人名
	ns	地名
	nsf	音译地名
	nt	机构团体名
	nz	其他专名
	nl	名词性惯用语
	ng	名词性语素
动词	v	动词
	vd	副动词
	vn	名动词
	vshi	动词“是”
	vyou	动词“有”
	vf	趋向动词
	vx	形式动词
	vi	不及物动词
	vl	动词性惯用语
	vg	动词性语素
形容词	a	形容词
	ad	副形词
	an	名形词
	ag	形容词性语素
	al	形容词性惯用语

在对分词后的词语进行极性统计时，本文还结合了 ICTCLAS 的分词特点选取了常用的 17 个否定词（表 3-4）。否定词对于微博文本的极性起到非常关键的作用，忽略否定词会对极性特征提取造成巨大的差错，可能得到截然相反的极性结果。因此，在判断情感词之后还应该检验该情感词之前是否包含了否定词，从而进行准确的极性划分。

表 3-4 常用否定词表

不能、不行、不曾、不必、不够、难以、终止、停止、放弃、反对、没有、少、不、别、未、反、没

3.2.3 情感特征分析

情感特征不仅要能够表示表示文本的内容，而且要具备与其他文本的区分能力。考虑微博的自身特点，本文将微博情感特征划分成以下几个方面：

1) 正负词和极性强度特征：通过对微博进行分词，结合情感词典和否定词表，统计微博的正负极性情感词个数，并统计极性强度的最大值。
2) 微博表情特征：结合微博表情极性分类表，统计微博中正面、负面、中性表情的数量。
3) 词性特征：词性特征是文本分类的重要特征，主观性文本中往往包含较多的形容词、副词、动词，而客观性文本一般包含较多名词，因此将各个词性的频率作为极性分类的部分特征。
4) 标点符号特征：在主观性文本中常常出现问号或感叹号，而连续的问号或感叹号往往表示着强烈的情感，因此将问号、感叹号、连续问号或感叹号作为极性分类的部分特征。

综上所述，本文将提取出如下 13 个极性特征：

表 3-5 微博情感特征集

特征	类型
正面词数量 poswn	int
负面词数量 negwn	int
正面词最大极性强度 posmax	int
负面词最大极性强度 negmax	int
正面表情数量 posfn	int
负面表情数量 negfn	int
中性表情数量 objfn	int
形容词数量 an	int
名词数量 nn	int
动词数量 vn	int
感叹号 hasExclamationMark	Boolean
问号 hasQuestionMark	Boolean
连续标点符号 hasContinuousMarks	Boolean

3.2.4 基于情感词典的微博情感特征提取算法

基于 3.2 小节的分析，我们提出了基于词典的微博情感特征提取算法：首先分析整理出情感词典、表情词典和否定词词典，写入数据库；接着对原始微博进

行链接、标签、@等媒体信息过滤，利用 ICTCLAS 分词辅助工具进行分词后通过情感词典、表情词典和否定词词典提取出极性词特征以及表情特征，并统计出形容词、名词、动词数量等词性特征和标点符号特征。具体算法流程如下：

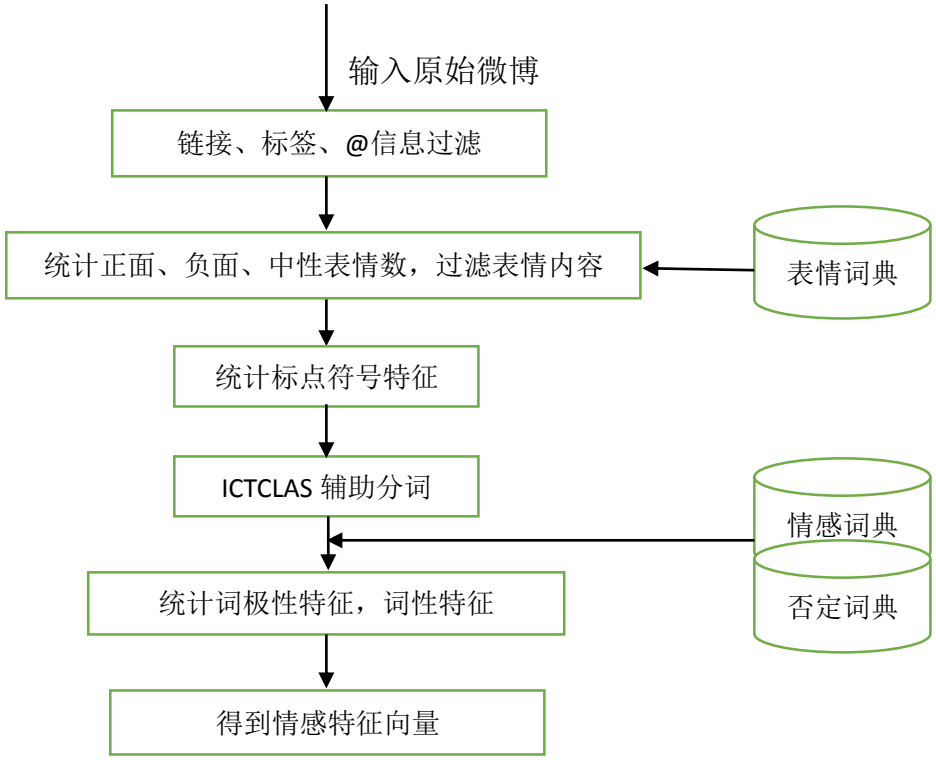

图 3-2 基于情感词典微博情感特征提取流程

情感特征提取算法伪代码如下：

Input：原始微博 MicroblogFile，情感词典 sentiment word dictionary，表情词典 face dictionary，否定词典 privative dictionary

Output：每条微博对应的情感特征向量 Fi

Extract positiveMap(Hashmap), negativeMap(Hashmap) from sentiment word dictionary; //提取出情感词和对应的情感强度

Extract positiveFace(Hashset), negativeFace(Hashset), objectiveFace(Hashset) from face dictionary; //提取出正面、负面、中性表情集合

Extract provative(Hashset) from privative dictionary; //提取出否定词集合

For Mi in MicroblogFile:

```
        Content = ChainingFilter(Mi);       //过滤链接
        Content = HashtagFilter(Content);       //过滤标签
        Content = MentionFilter(Content);       //过滤@信息
        Fi = initialize(poswn, negfn, posmax, negmax, posfn, negfn, objfn,…
an, nn, vn, hasExclamationMark, hasQuestionMark, hasContinuousMarks);
        Fi = extract(Content, positiveMap, negativeMap, positiveFace,…
negativeFace, objectiveFace, provative);       //提取特征
End
```

3.3 基于 SVM 的微博情感极性分类

本文的微博情感特征虽然都为离散值，但是由于样本差异大，某些特征值取值范围大，且分布十分不均，因此在分类之前针对具体分布情况对情感特征值进行了分段处理，具体分析在 4.3.1 和 4.3.2 小节实验分析部分。

由于本文中把微博情感分为正面、负面、中性三类，因此传统的二分类 SVM 无法直接使用，需要使用多个标准的二分类 SVM。常见的方法有 one-against-one 和 one-against-all 两种方法。One-against-one 在两两样本之间设计一个 SVM，对于 k 个类别的样本，一共需要 k(k-1)/2 个 SVM。分类时将测试样本放入所有 SVM 中进行分类，采取投票形式，最终划分为得票最高的那一类。本文中使用的 libsvm 就是采用 One-against-one。

3.3.1 SVM 参数优化算法

在第二章的有监督学习算法详述小节介绍了基本 SVM 的数学理论，但是有时候为了更好地拟合或在面对样本特征线性不可分的情况时，我们必须使用核函数对原始数据进行空间映射。SVM 中常用的核函数有径向基核函数、线性核函数、多项式核函数。其中应用最广泛的是径向基核，它可以将样本映射到无穷维空间。

1) 径向基核（radial basis）：$K(x, y) = \exp(-\gamma||x - y||^2)$ (3.1)
2) 线性核（linear）：$K(x, y) = x^T y$ (3.2)
3) 多项式核（polynomial）：$K(x, y) = (\gamma x^T y + c)^d$ (3.3)

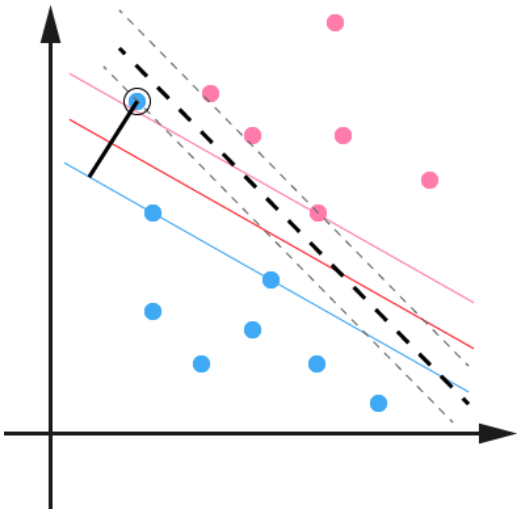

图 3-3　离群点 outlier 对 SVM 的影响

核函数对线性不可分数据映射到高维空间后，虽然线性分隔的概率大大增加，但是还有可能因为数据噪音（偏离正常位置很远的数据点，离群点 outlier）的存在而影响最优间隔超平面。如图 3-3 所示，黑色圆圈中的蓝点较大偏离了蓝色区域导致超平面 margin 变小，如果它继续向右上方偏移的话我们将无法构造出分割超平面。为了处理这种情况 SVM 引入了松弛变量ε（slack variable），允许数据点一定程度上偏离超平面。这样 SVM 的最优化目标函数就变为：

$$\begin{aligned} & min_{\gamma,\mathrm{w},\mathrm{b}} \ \frac{1}{2}||\mathrm{w}||^2 + C\sum_{i=1}^{n}\varepsilon_i \\ & \text{s.t.} \ \ y^{(i)}\left(\mathrm{w}^T x^{(i)} + b\right) \geq 1 - \varepsilon_i, \ \mathrm{i} = 1, \ldots, \mathrm{n} \\ & \varepsilon_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{3.4}$$

公式 3.4 中 C 是事先确定好的常量，为最优化函数中对离群点的惩罚因子，用于平衡“寻找最大超平面“和“保证数据点偏差量最小”。

选取径向基核函数后需要对参数 C 和径向基核函数（公式 3.1）中的γ进行优化。C 为最优化函数中对离群点的惩罚因子，γ决定了数据映射到新的特征空间后的分布。为了找到合适的 C 和γ，同时解决过拟合问题，我们首先需要通过交叉验证（cross validation）对一定范围内的 C 和γ进行网格搜索（grid search）[13]，然后选取交叉验证正确率最高的参数对（C，γ）作为训练模型的参数值。V-折交叉验证是先将训练样本分成 V 个相同的子集，之后顺序地对 V-1 个子集进行模型训练，余下的 1 个子集作为验证，最后得到平均准确率。

交叉验证的网格搜索算法对参数 C 和γ的优化过程如下：

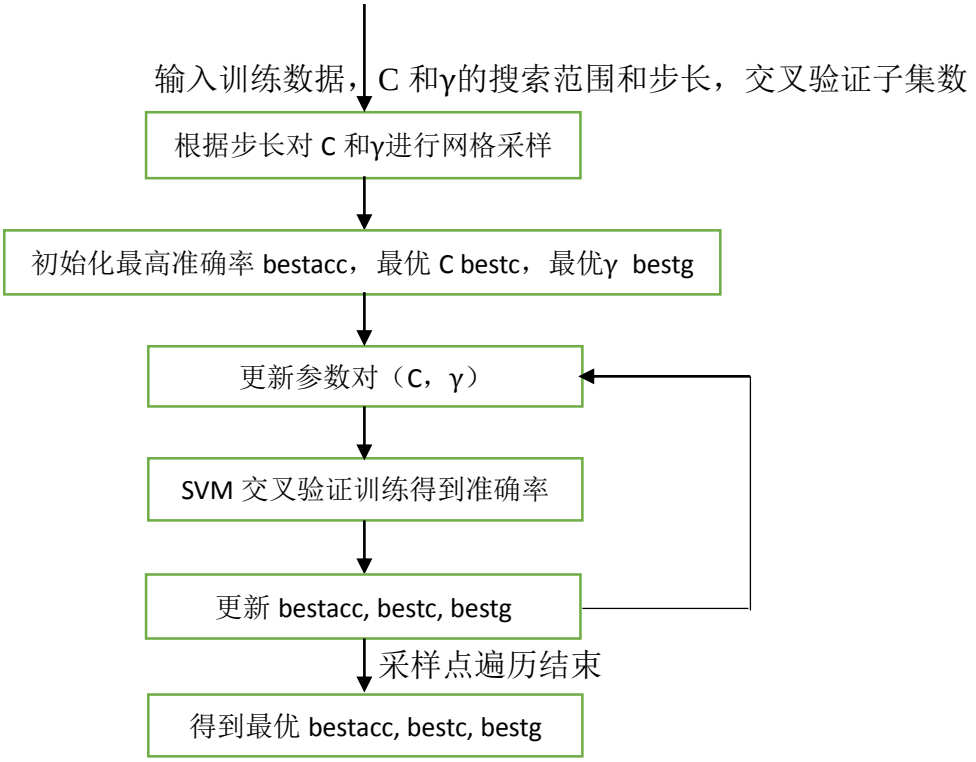

图 3-4 交叉验证的网格搜索算法对参数 C 和γ的优化过程

3.3.2 基于 SVM 的微博极性分类算法

本文对情感特征预处理后（4.3.1 和 4.3.2 小节），利用参数优化的 SVM 对微博进行情感极性预测，随机取 70%作为训练数据，30%作为测试数据，训练时采用 5 折交叉验证，训练预测过程如下：

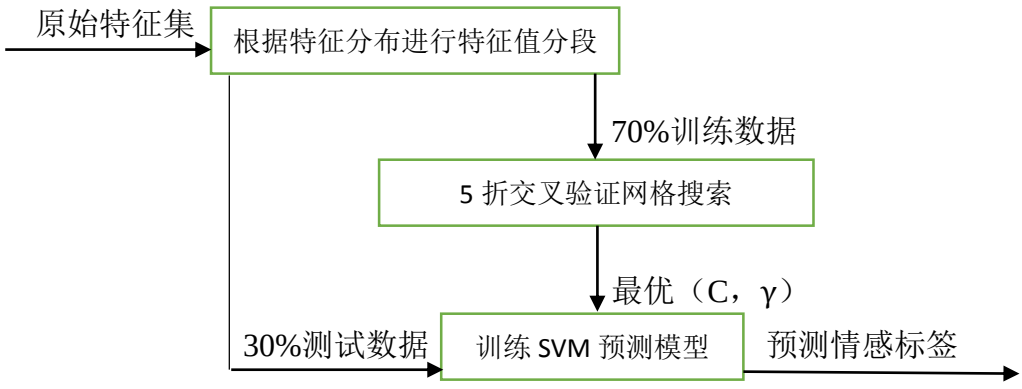

图 3-5 基于 SVM 的微博极性分类流程

支持向量机分类的算法如下（以 libsvm 为工具）：

```
Input：情感特征矩阵（每一行都为情感特征向量，第 1 列为标签 label，2 至 14 列分别为 13 个特征值 feature）
Output：测试集预测结果 predict_label
//对情感特征矩阵随机取 70%当作训练数据，30%当作测试数据
[randomLabel, randomFeature] = reorder(label, feature);
train = randomFeature (1:floor(2*end/3),:);
train_label = randomLabel (1:floor(2*end/3),:);
test = randomFeature (floor(2*end/3)+1:end,:);
test_label = randomLabel (floor(2*end/3)+1:end,:);
//采用 cross-validation 和 grid-search 优化参数 C 和γ
[bestaccuracy, bestc, bestg] = SVMcgForClass(train_label, train);
//训练模型
model = svmtrain(train_label, train, bestc, bestg);
//预测
[predict_label, accuracy, decision_values] = svmpredict(test_label, test, model)
```

3.4 信息增益分析微博情感特征影响力

由于本文的情感特征都为离散值，在对不平衡特征重新分段后，通过计算各个情感特征的信息增益（Information Gain），根据信息增益大小来判断特征对分类的重要程度。

3.4.1 信息量和信息熵

在日常生活中，司空见惯的事件往往不太引人注意，而极少发生的事件一旦发生则较容易吸引人们的关注。换句话说，小概率事件包含的信息量越大。假设有一值域为$\{x_1, x_2, \ldots x_n\}$的随机变量 X，每个值x_i发生的概率为$\mathrm{p}(x_i)$，则x_i的信息量表示为：

$$\mathrm{I}(x_i) = -log_2(\mathrm{p}(x_i)) \tag{3.5}$$

以汉字和英文字母为例：为了简单说明问题，假设每个汉字和英文字母在文章中出现的概率相同，常用的汉字有 2500 个，每个汉字的信息量为 11.3；而英文字母有 26 个，每个字母的信息量为 4.7。这意味着汉字包含的信息量高于英文

字母，写同一篇文章用汉字比英文更为简短。

信息熵为信息量的期望，表示对系统不确定性的测量。对于以上例子，熵的具体定义如下：

$$\mathrm{H(X)} = \mathrm{E[I}(X)] = \sum_{\mathrm{i}} \mathrm{p}(x_i)\, \mathrm{I}(\mathrm{x}_i) = -\sum_i \mathrm{p}(x_i)\, log_2(\mathrm{p}(x_i)) \quad (3.6)$$

当随机变量 X 的分布较为均衡时，X 的熵较高，不确定性大，可预测性更低；当 X 分布不均衡时，X 的熵较低，不确定性小，可预测性更高。

3.4.2 微博情感特征信息增益

信息增益[19]是一种衡量样本特征重要性的方法。对于一个分类系统来说，特征 A 的信息增益表示特征能够为分类带来多少信息，带来的信息越多那么特征 A 越重要。特征 A 的信息增益定义为类别的经验熵 H(C)和给定特征 A 时类别的条件熵 H(C|A)之差，具体表示为：

$$\mathrm{IG(A)} = \mathrm{H(C)} - \mathrm{H(C|A)} \quad (3.7)$$

本文的情感特征集包含了 13 个情感特征，将微博分为正面 p、负面 n 和中性 o 三类，因此本文类别包含的信息熵（经验熵）为：

$$\mathrm{H(C)} = -\sum_{i=1}^{3} \mathrm{p}(C_i) log_2(\mathrm{p}(C_i)) \quad (3.8)$$

假设特征 A 的取值为$\{A_1, A_2, \dots A_n\}$，条件熵 H(C|A)的计算公式为：

$$\mathrm{H(C|A)} = \sum_{i=1}^{n} \mathrm{p}(A_i)\, H(C|A_i) = -\sum_{i=1}^{n} (\mathrm{p}(A_i) \sum_{j=1}^{3} \mathrm{p}(C_j|A_i) log_2(\mathrm{p}(C_j|A_i))) \quad (3.9)$$

通过对正面词数量、负面词数量、正面词最大极性强度、负面词最大极性强度、正面表情数量、负面表情数量、中性表情数量、形容词数量、名词数量、动词数量、是否包含感叹号、是否包含问号、是否包含连续标点符号 13 个情感特征分别计算信息增益，比较相对大小可以得出特征对于情感分类的重要程度。

3.5 本章小结

本章提出了本文的研究和实验方案。针对微博情感分析详述了本文微博情感特征的提取方法（情感词典，情感表情、媒体特征预处理、分词、特征提取算法）；详述了支持向量机的原理、参数优化、训练测试流程；详述了通过计算信息增益来衡量本文情感特征的重要程度。

第4章 实验与分析

4.1 实验环境与实验数据

情感分析实验的数据处理和特征提取部分是在 Eclipse 平台上使用 Java 实现的，特征集的训练测试部分使用 Matlab 进行实现。

4.1.1 微博 API

为了便于开发应用，新浪微博为开发者提供了一套完整的 API 接口用于获取用户资料、微博内容等社交信息，为用户提供个性化服务。

Oauth 是一种国际通用的授权方式，调用新浪微博 API 首先需要进行 OAuth 认证，相对 OAuth1.0，OAuth2.0 的授权流程更加简单安全，是最主要的用户身份验证和授权方式。OAuth2.0 的授权步骤如图 4-1 所示（其中 Client 为第三方应用，Resource Owner 为用户，Authorization Server 为授权服务器，Resource Server 为 API 服务器）：

图 4-1 OAuth2.0 认证步骤

为了获取新浪 API 的调用权限，必须进行以下操作步骤：

- 首先在 API 开放平台上申请一个第三方应用（移动应用或网站接入），获取 client_id 和 client_secret，同时填写授权码回调地址（redirect_uri）。
- 然后将 client_id 和回调地址打包发送至新浪微博 API 进行授权请求，获得回

调地址中返回的授权码（code），发送方式如下：

（https://api.weibo.com/oauth2/authorize?client_id=YOUR_CLIENT_ID&response_type=code&redirect_uri=YOUR_REGISTERED_REDIRECT_URI）

- 最后将 client_id、client_secret、回调地址、获得的授权码发送至 API 验证并请求接入令牌（access token），发送方式如下：

（https://api.weibo.com/oauth2/access_token?client_id=YOUR_CLIENT_ID&client_secret=YOUR_CLIENT_SECRET&grant_type=authorization_code&redirect_uri=YOUR_REGISTERED_REDIRECT_URI&code=CODE）

- 获得接入令牌（access token）之后就可以利用 Java SDK 等开发工具包进行微博 API 的调用。其中搜索话题微博 API 为 search/topics，对返回的 JSON 格式数据进行解析，提取出微博内容。

4.1.2 实验数据

本实验从新浪微博中选取了转基因、雾霾、房价三个话题中情感倾向较明显、篇幅较长的共 300 条微博进行人工标注，分为正面、负面、中性三类，标注结果如下表所示：

表 4-1 人工标注语料集

话题	正面	负面	中性	总和
转基因	31	51	18	100
雾霾	40	30	30	100
房价	25	45	30	100
总和	96	126	78	300

本文还将算法应用在 NLP&CC 2012 微博情感语料集中（将微博拆分为单句，不含表情特征，XML 格式），其中包含了 20 个话题共 3275 条微博，其中正面微博 433 条，负面微博 1687 条，中性微博 1155 条，如下表所示：

表 4-2 NLP&CC2012 微博情感语料集

话题	正面	负面	中性	总和
菲军舰恶意撞击	5	121	94	220
疯狂的大葱	6	76	63	145
官员财产公示	16	115	38	169

续表 4-2

话题	正面	负面	中性	总和
官员调研	20	99	46	165
国旗下讨伐教育制度	59	49	55	163
韩寒方舟子之争	21	111	90	222
假和尚搂女子	8	107	32	147
奖状植入广告	23	53	57	133
90 后暴打老人	4	94	59	157
90 后当教授	110	13	12	135
六六叫板小三	28	105	84	217
名古屋市长否认南京大屠杀	1	56	65	122
彭宇承认撞了南京老太	11	80	65	156
皮鞋果冻	2	110	35	147
苹果封杀 360	57	58	41	156
三亚春节宰客	12	110	79	201
食用油涨价	8	70	45	123
洗碗工留剩菜被开除	7	57	37	101
学雷锋被钓鱼执法	9	94	80	183
中国教师收入全球几垫底	26	109	78	213
总和	433	1687	1155	3275

4.2 评价指标

对于正面、负面和中性微博每一类的分类评价，我们把预测实例分为正类和负类。以正面微博为例，正类（1）为正面微博，负类（0）为负面微博或中性微博，表 4-3 说明了正面微博的分类情况，负面微博和中性微博同理。

表 4-3 正面微博 TP、FN、FP、TN 定义

		预测	
		1（正面）	0（负面或中性）
实际	1（正面）	True Positive(TP)	False Negative(FN)
	0（负面或中性）	False Positive(FP)	True Negative(TN)

对于正面微博来说，精确率 p 的计算公式为$p = \frac{TP}{TP+FP}$，表示被预测成正面微博中正面微博所占的比例；召回率 r 的计算公式为$r = \frac{TP}{TP+FN}$，表示正面微博中被预测为正面微博所占的比例。

精确率和召回率的评价角度不同且相互影响，所以引入 F 值对精确率和召回率进行综合评价。F 值的计算公式如下：

$$F = \frac{2pr}{p+r} = \frac{2TP}{2TP+FP+FN} \tag{4.1}$$

4.3 实验结果与分析

4.3.1 人工标注集特征分布

经过对原始微博进行媒体特征过滤，分词，极性词统计，极性强度统计，情感表情符号统计，情感标点符号统计后，每条微博得到了正面词个数、负面词个数、正面词极性强度、负面词极性强度、正面表情数、负面表情数、中性表情数、形容词数、名次数、动词数、是否包含问号、是否包含感叹号、是否包含连续符号共 13 个情感特征，其中特征分布如下所示：

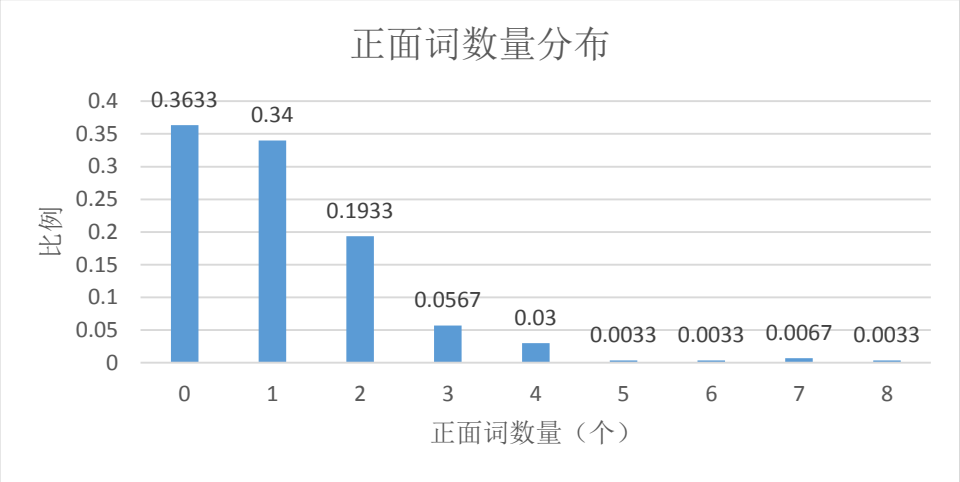

图 4-2 正面词数量分布

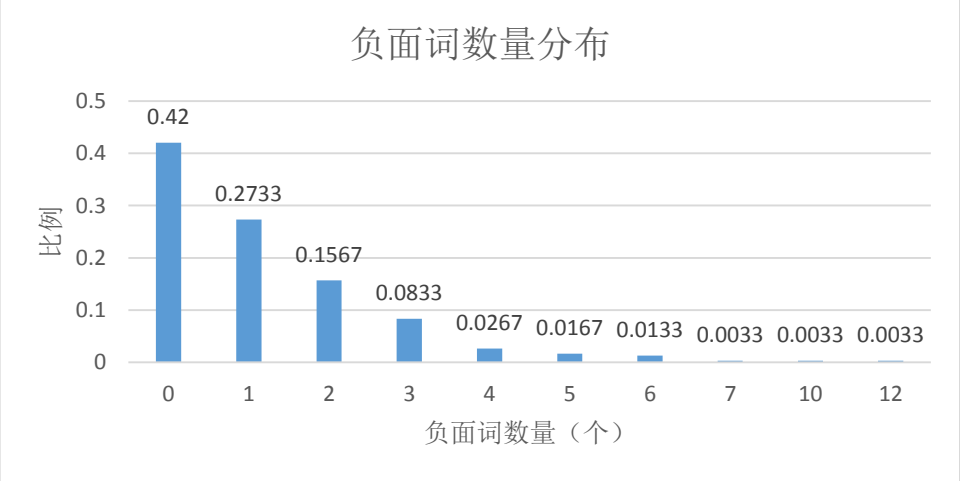

图 4-3 负面词数量分布

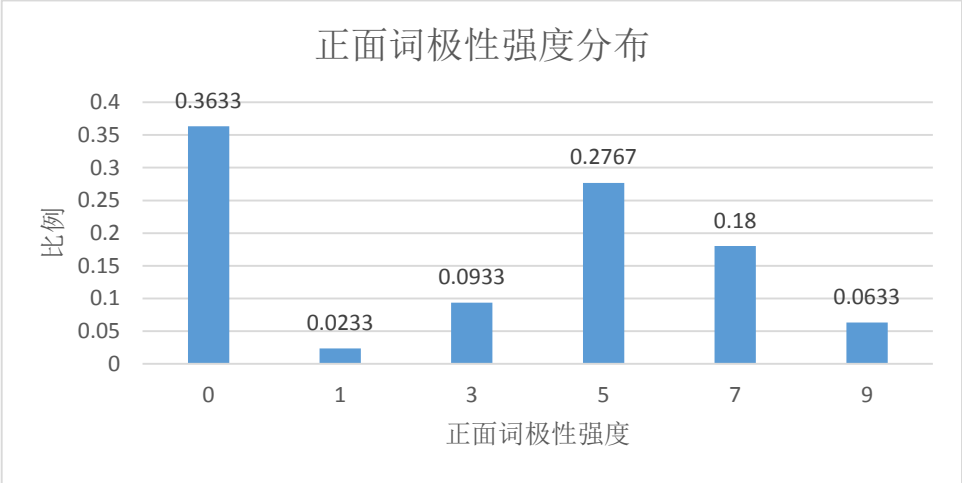

图 4-4 正面词极性强度分布

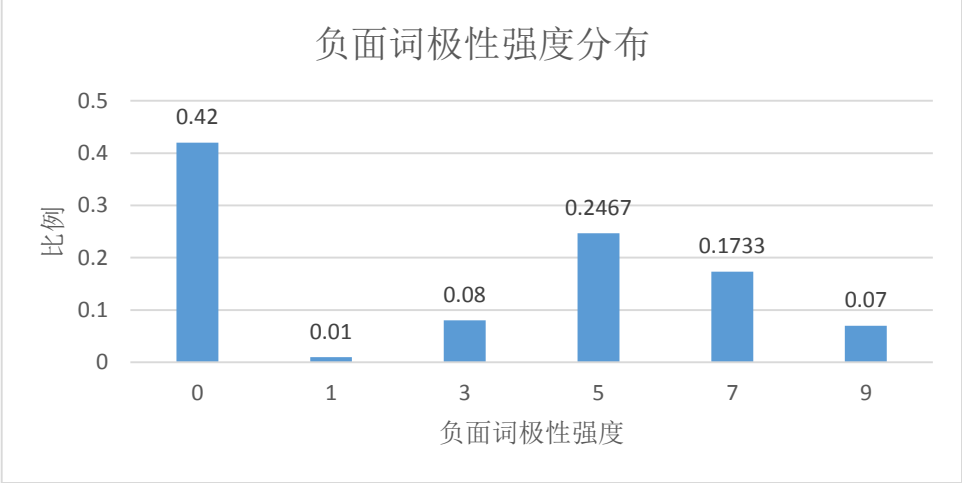

图 4-5 负面词极性强度分布

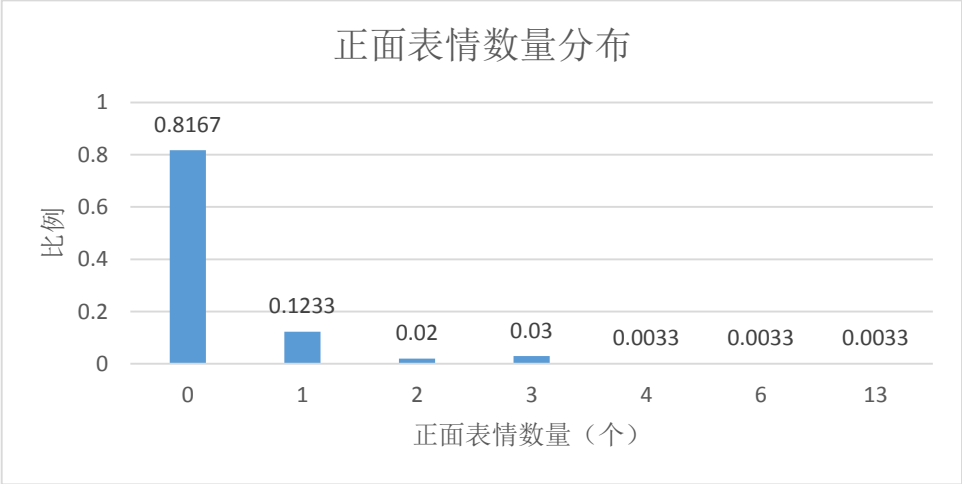

图 4-6 正面表情数量分布

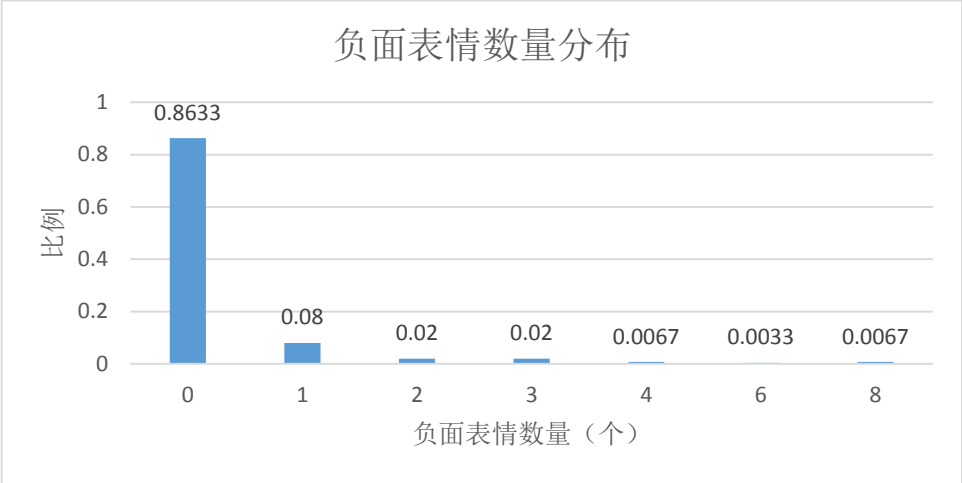

图 4-7 负面表情数量分布

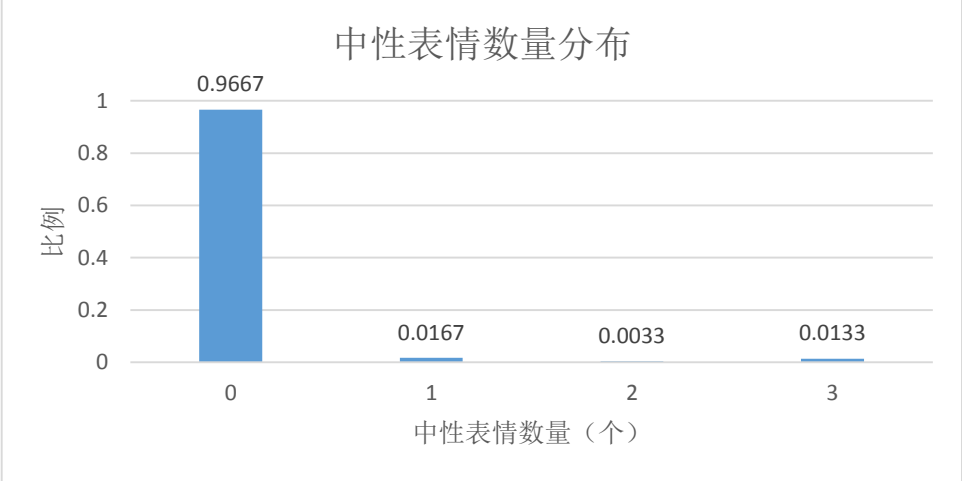

图 4-8 中性表情数量分布

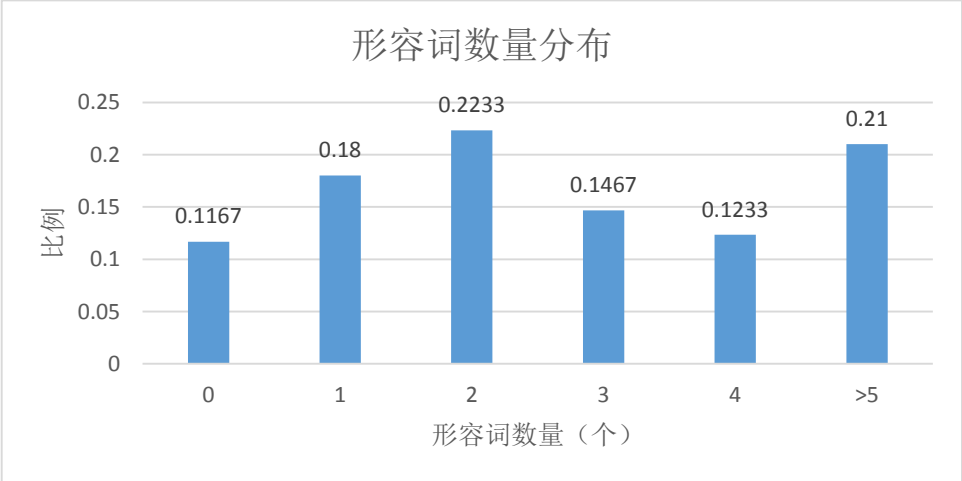

图 4-9 形容词数量分布

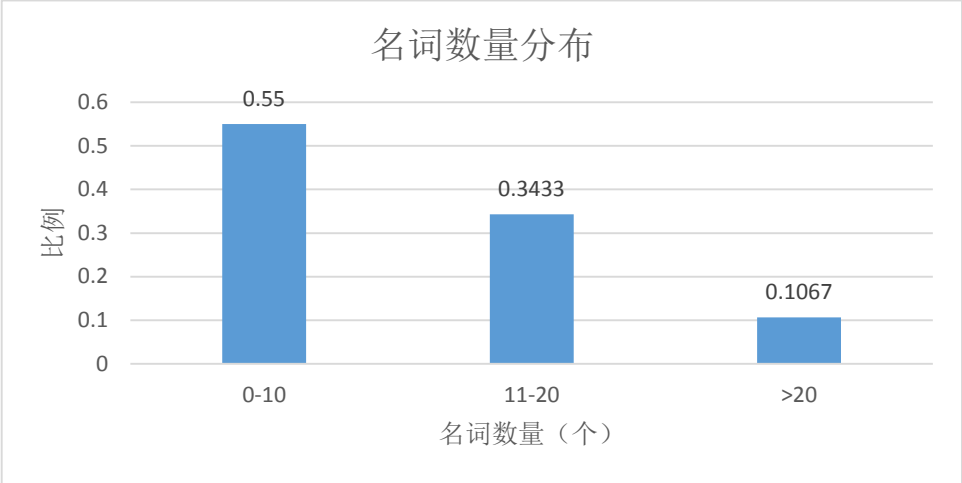

图 4-10 名词数量分布

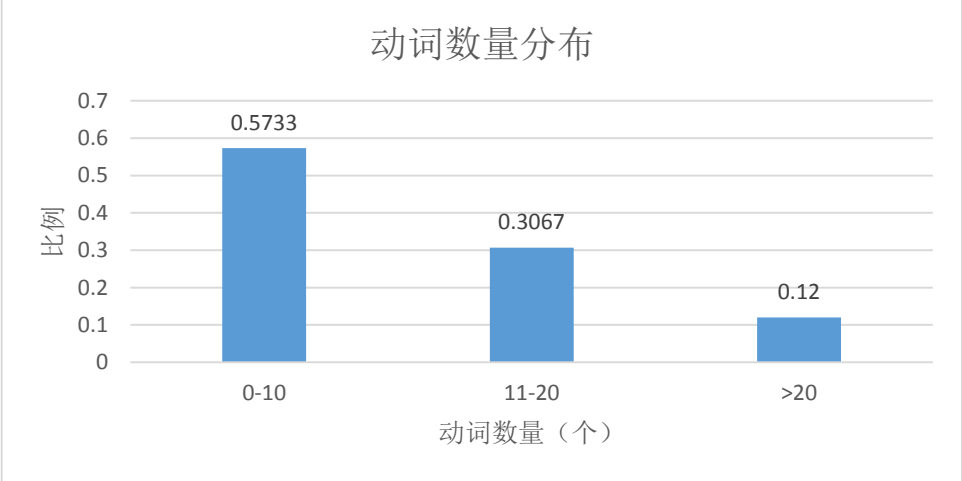

图 4-11 动词数量分布

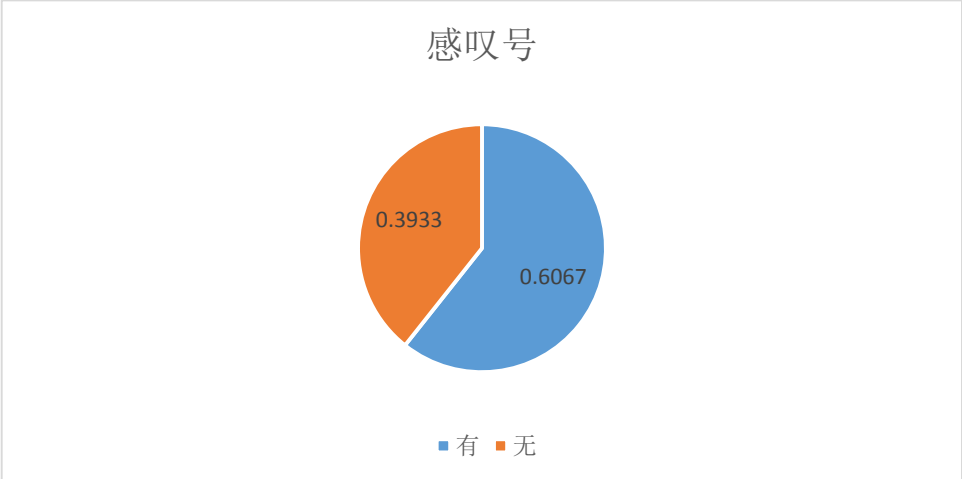

图 4-12 是否包含感叹号

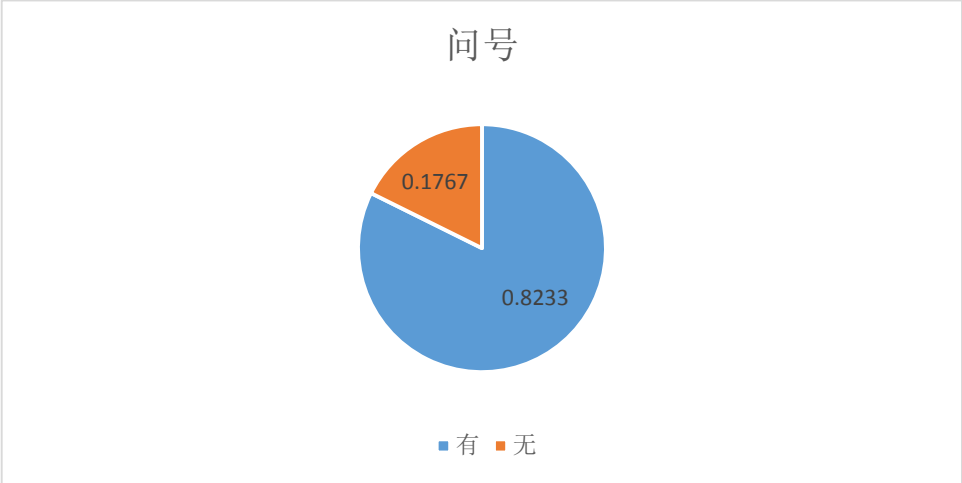

图 4-13 是否包含问号

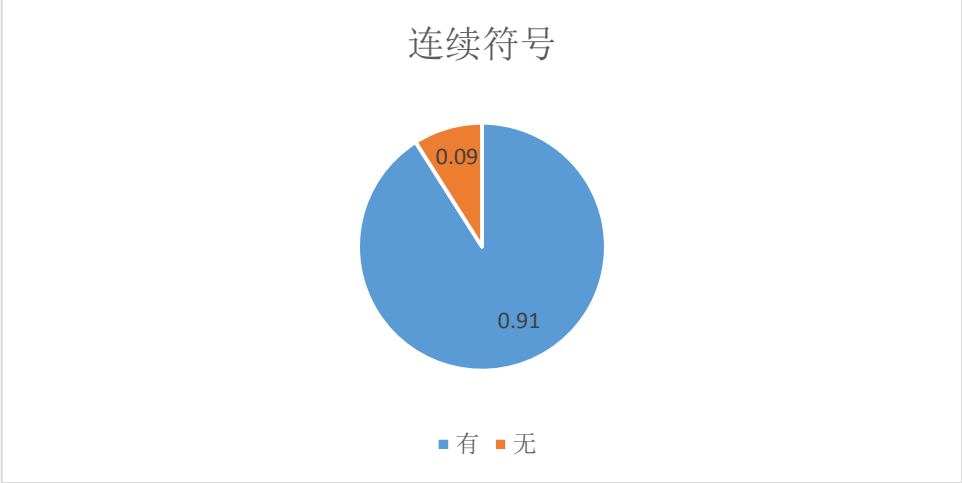

图 4-14 是否包含连续符号

通过观察提取出的离散特征的分布，我们发现正面词和负面词的数量分布普遍集中在 0,1,2,3，而词的最大极性强度则集中在 0,5,7 三类，带表情的特征和标点符号特征相对较少，而形容词、名词、动词各词性的分布则更加离散。

4.3.2 原始特征预处理

由于原始情感特征集分布非常不平衡，因此本文对提取出的原始特征进行了预处理，根据各特征的离散值的概率分布进行特征值分段来加快分类速度，提高分类准确率。

其中，“正面词数量”分为 0,1,2,>=3 四种情况，“负面词数量”分为 0,1,2,>=3 四种情况，“正面表情数量”分为 0,1,>=2 三种情况，“负面表情数量”分为 0,1,>=2 三种情况，“中性表情数量”分为 0,>=1 两种情况，“形容词数量”分为 0,1,2,3,4,>=5

五种情况，“名词数量”分成 0-10,11-20,>20 三种情况，“动词数量”分为 0-10,11-20,>20 三种情况。情感特征预处理前后 SVM 分类效果比较：

表 4-4 原始特征预处理前后 SVM 测试结果对比(人工标注集)

算法	极性	Precision%	Recall%	F measure%	Support
预处理前	中性	74.2	57.5	63.8	27.4
	正面	81.5	79.9	80.5	32.0
	负面	76.5	88.4	81.7	40.6
	加权平均	77.8	77.0	76.5	
	最大	80.0	79.0	79.4	
预处理后	中性	75.2	65.1	69.2	26.4
	正面	83.4	89.9	86.4	32.0
	负面	84.3	84.9	84.4	41.6
	加权平均	81.6	81.4	81.1	
	最大	83.8	84.0	83.9	

从表 4-4 可以观察到在对不平衡特征集进行分段处理之后，SVM 的预测效果有了明显的提升（平均精确度 81.6>77.8；平均召回率 81.4>77.0；平均 F 值 81.1>76.5）。

4.3.3 人工标注集预测结果与分析

训练过程中，将数据随机分为 70%的训练数据和 30%的测试数据，并且运行 30 次算法取平均值。

支持向量机、朴素贝叶斯、K 近邻和随机森林对中性、正面和负面微博的预测结果如下（均四舍五入取到小数点后一位）：

表 4-5 测试结果(人工标注集)

算法	极性	Precision%	Recall%	F measure%	Support
KNN (K=4)	中性	60.6	56.6	57.7	25.5
	正面	72.4	76.9	74.1	31.5
	负面	74.0	73.2	73.2	43.0
NB	中性	70.3	66.1	67.4	26.2
	正面	80.7	85.7	82.9	31.4
	负面	80.0	79.1	79.1	42.4
RF (treenum=500)	中性	69.8	70.3	69.5	25.4
	正面	84.5	85.3	84.5	32.8
	负面	81.5	80.1	80.6	41.8

续表 4-5

算法	极性	Precision%	Recall%	F measure%	Support
SVM（高斯核，5 折交叉验证）	中性	75.2	65.1	69.2	26.4
	正面	83.4	89.9	86.4	32.0
	负面	80.4	83.8	81.9	42.8

观察 SVM、NB、KNN 和 RF 算法在各个分类上的测试结果，我们可以发现各算法对正面微博和负面微博的预测精度要明显高于对中性微博的预测精度。由于中性微博介于正面和负面之间，且相对情感微博而言，中性微博的特征并不那么明显，且训练量小，所以预测精度较差能够理解。四种算法对于正面微博的预测精度略高于负面微博（从正面微博和负面微博的 F 值来看，SVM 的 86.4 对 81.9；RF 的 84.5 对 80.6；NB 的 82.9 对 79.1；KNN 的 74.1 对 73.2）。

表 4-6 和图 4-15 展示了四种算法在样本总体上的 precision、recall 和 F 值。

表 4-6 四种算法最终分类结果（人工标注集）

	Precision%	Recall%	F measure%
SVM	81.6	81.4	81.1
NB	78.1	77.4	77.2
KNN	70.6	69.8	69.6
RF	79.8	79.2	79.2

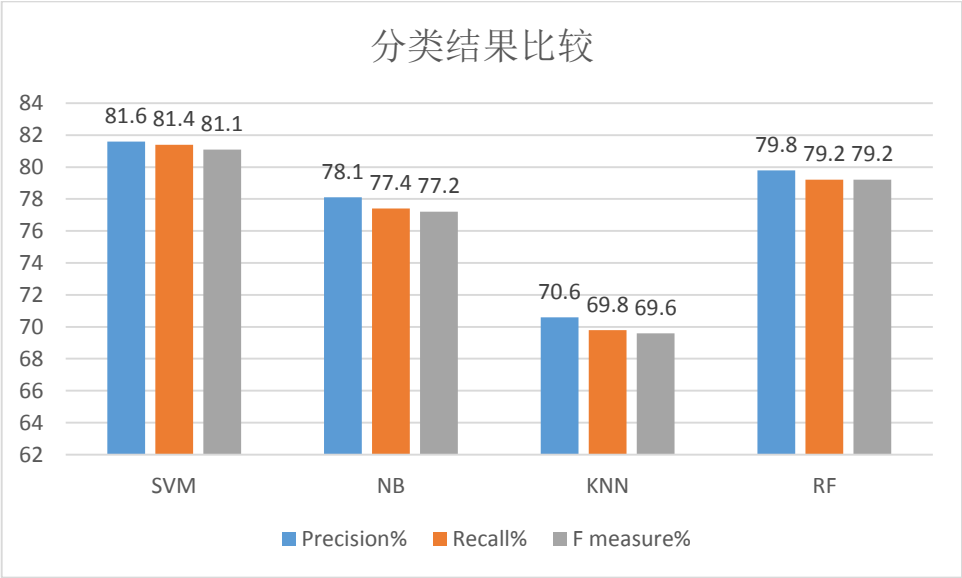

图 4-15 分类结果比较（人工标注集）

对比四个算法在整体预测上的效果，我们可以发现对于人工标注数据集，

SVM 在 precision、recall 和 F-score 方面的预测精度都略高于 RF 和 NB，KNN 的预测效果明显比较差。

4.3.4 NLP&CC 2012 微博情感标注集预测结果与分析

四种算法在 NLP 语料集上中性、正面和负面微博的测试结果如下：

表 4-7 测试结果（NLP 语料集）

算法	极性	Precision%	Recall%	F measure%	Support
NB	中性	48.4	56.9	52.2	389.2
	正面	38.4	45.6	41.6	145.6
	负面	67.3	55.6	60.8	557.2
KNN	中性	51.7	44.0	47.4	389.3
	正面	44.0	25.3	31.8	140.9
	负面	61.3	74.2	67.1	561.8
RF	中性	50.9	38.9	43.8	391.7
	正面	43.9	24.5	31.2	148.5
	负面	59.4	76.3	66.8	551.8
SVM	中性	57.3	37.0	44.5	384.0
	正面	48.5	27.0	34.7	142.6
	负面	60.6	81.6	69.5	565.4

观察正面、负面、中性每一类的预测结果，我们发现四类算法对于负面微博的预测精度极大地高于正面微博（以 F 值为例，SVM 的 69.5 对 34.7；RF 的 66.8 对 31.2；KNN 的 67.1 对 31.8；NB 的 60.8 对 41.6），一个主要原因是负面微博的样本量占了一半以上（1687），训练充分，而正面微博样本量占了八分之一（433），训练不够充分；另一个原因可能是 NLP 语料集中正面微博的倾向性不是太明显，中文情感词汇本体库中的正面词汇集不足以表达正面微博。

表 4-8 展示了四种算法在 NLP 样本总体上的 precision、recall 和 F 值。

表 4-8 四种算法最终分类结果

	Precision%	Recall%	F measure%
SVM	57.9	58.8	56.1
NB	56.8	54.7	55.2
KNN	55.7	57.1	55.5
RF	54.3	55.8	53.7

观察表 4-8 实验结果我们发现参数优化的 SVM 在 NLP 语料集上的预测效果相对随机森林、朴素贝叶斯和 K 近邻要好。

另外，相对人工标注集，我们发现 NLP 数据集的模型训练效果显著降低，可能有如下原因：相比前文的人工标注集，NLP 语料集把一条完整微博分成了一个个单句，因此篇幅短，某些句子情感倾向不太明显，提取出的特征非常稀疏，且不含表情特征；人工标注集是对完整的一条微博进行特征提取，包含多个句子，篇幅较长，在筛选时选择的微博情感倾向性较明显，且包含表情特征，提取出的特征区分性高；另外，NLP 语料集的样本分布十分不平衡，从表 4-7 的测试结果可以明显观察到负面微博的精确度显著高于正面微博的精确度,因此本文基于情感词典的有监督学习方法对于 NLP 语料集不太适用。

4.3.5 信息增益分析特征影响力

对人工标注集的 13 个情感特征分别计算信息增益，结果如下：

表 4-9 情感特征的信息增益（人工标注集）

特征序号	特征	信息增益(bit)
1	正面词数量	0.2437
2	负面词数量	0.2477
3	正面词最大极性强度	0.2107
4	负面词最大极性强度	0.2758
5	正面表情数量	0.2785
6	负面表情数量	0.0577
7	中性表情数量	0.0088
8	形容词数量	0.0147
9	名词数量	0.0401
10	动词数量	0.0637
11	感叹号	0.1408
12	问号	0.0692
13	连续标点符号	0.0332

从表 4-9 可以发现“正面表情数量”的信息增益最大（0.2785），可以认为其在人工标注集上的影响力最大。另外情感词的极性特征总体来说有着更大的信息增益（“正面词数量”=0.2437，“负面词数量”=0.2477，“正面词最大极性强度”=0.2107，“负面词最大极性强度”=0.2758），即情感词对情感分类的作用更大，这也符合我们直观的判断。

另外我们观察到“正面表情数量”的信息增益（0.2785）比“负面表情数量”的信息增益（0.0577）大很多，按照常理来说两者应该相差不太大。但是注意到信息增益受到数据集分布的影响，一般分布越均匀，信息增益越大。观察图 4-6、图 4-7 和图 4-8 的正面、负面、中性表情分布情况，虽然带表情的微博很少，但还是可以看出“正面表情数量”相对“负面表情数量”和“中性表情数量”的分布相对更加均匀，因此“正面表情数量”的信息增益较大可以理解。

4.4 本章小结

本章主要利用第三章中的特征提取方法对公开语料集和手工标注集提取相应情感特征向量，根据特征分布进行预处理并利用 SVM、NB、KNN、RF 四种有监督学习分类算法对特征集进行训练测试优化，对比不同情感分类模型的预测结果并对结果进行了分析，最后利用信息增益衡量了各个特征对情感分类的重要性。

第5章 本文总结

5.1 论文主要工作

本文查阅了情感分析的国内外研究现状，学习了情感分析的研究方法。

针对微博情感分析制定了研究方案，提出了基于情感词典和 SVM 的预测模型。获取了新浪微博中关于转基因、雾霾和房价三个话题的微博并且进行有选择的人工标注，将微博内容分为正面、负面和中性三个类别；分析了微博内容中的情感因素，然后利用分词工具、情感词典、表情词典对微博内容进行了预处理和情感特征的提取。

之后对提取出的不平衡情感特征集进行基于概率分布的分段预处理，深入学习了支持向量机的原理和参数优化，并和随机森林、朴素贝叶斯、K 近邻算法的预测结果进行对比，发现 SVM 的效果最好。同时，本文还使用了 NLP&CC 2012 微博情感语料集进行有监督学习的预测，与人工标注语料集进行了对比，分析了差别和原因。

最后，本文计算了特征集中各个情感特征的信息增益，衡量了各特征对于本文微博情感分类的重要程度。

5.2 将来的工作

在特征提取方面，本实验采用的是中科院 ICTCLAS 中文分词系统与大连理工大学中文情感本体库相结合的情感词频率和情感强度特征提取。虽然 ICTCLAS 对中文分词的精确度较高，但是微博的语法一般不太规范，网络用语较多，对分词的效果有所影响，在之后的研究中要针对网络用语进行特殊处理；另外中文情感本体库虽然词库量大，较为完整，但是在分词的基础上，有些较为复杂的情感词可能无法匹配到，也会对特征提取有所影响；本文虽然针对 ICTCLAS 分词系统进行了否定词的探测，但是对于结构更加复杂的否定句和双重否定句，无法进行有效地判断，在之后的学习中将进一步改进。

在分类算法方面，针对稀疏特征集，不平衡特征集的优化处理、训练参数的优化都有待进一步的研究。

参考文献

[1] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions[J]. Conference on Empirical Methods in Natural Language Processing (EMNLP-03, 2003:105-112.

[2] Hu M, Liu B. Mining and summarizing customer reviews[J]. in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004:168-177.

[3] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.

[4] Saif H, Fernandez M, He Y, et al. Senticircles for contextual and conceptual semantic sentiment analysis of twitter[M]//The Semantic Web: Trends and Challenges. Springer International Publishing, 2014: 83-98.

[5] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[6] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]//Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005: 625-631.

[7] Moraes R, Valiati J F, Neto W P G. Document-level sentiment classification: An empirical comparison between SVM and ANN[J]. Expert Systems with Applications, 2013, 40(2):621–633.

[8] 张华平. 基于多层隐马尔科夫模型的中文词法分析[C]//第 41 届 ACL 会议暨第二届 SIGHAN 研讨会, 日本. 2003: 63-70.

[9] 李正华, 车万翔, 刘挺. 基于 XML 的语言技术平台[J]. 第五届全国青年计算语言学研讨会论文集, 2010.

[10] 朱嫣岚, 闵锦, 周雅倩等. 基于 HowNet 的词汇语义倾向计算[J]. 中文

信息学报, 2006, (1):14-20.

[11] 徐琳宏，林鸿飞，潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2):180-185.

[12] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1-4.

[13] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[J]. 2003.

[14] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[15] Bishop C M. Pattern recognition and machine learning[M]. New York: springer, 2006: 205-206

[16] Bishop C M. Pattern recognition and machine learning[M]. New York: springer, 2006: 21-24

[17] Altman N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[18] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[19] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

致谢

首先，我衷心感谢我的导师郑小林副教授在学习和生活中给予我的谆谆教诲和悉心关怀。本论文从选题、构思、修改到成文，每个环节都凝结了实验室老师大量的心血，充斥了实验室学长学姐的耐心指导。郑小林教授专业知识渊博，治学态度严谨，在学术上精益求精、积极进取，在工作中求真务实，在生活中平易近人，给我留下了深刻的印象。在短短几个月相处时间里，我深深感受到了郑老师出色的人格魅力，这在以后的求学道路上无疑给我树立了旗帜鲜明的方向标，是我一生取之不尽的宝贵财富。

此外，我要感谢浙江大学电子服务研究中心的所有师兄师姐们，在学习和工作中离不开他们的支持和帮助。

感谢计算机学院的各位老师，谢谢你们在我本科生阶段带领我进入丰富多彩的计算机科学与技术的世界。

感谢对我本科论文进行评审的各位专家教授，感谢你们对论文的指导和提出的宝贵意见。

我要特别感谢我的父母，他们在我的成长中起到了不可磨灭的影响，是我人生中最重要的依靠和指引。

最后，感谢所有给予我指导、帮助、关心和支持的老师、亲人和朋友们。

林炜华

2015 年 6 月

本科生毕业论文（设计）任务书

一、题目：基于 SVM 的微博情感分析研究与实验

二、指导教师对毕业论文（设计）的进度安排及任务要求：

该毕业论文要求对情感分析相关国内外研究现状进行深入分析的基础上，提出情感分析预测模型，并重点围绕微博情感分析、分类算法展开研究。实验数据要求采用人工标注情感集和 NLP&CC 语料集，实验结果要清晰可信。

进度安排如下：

2015 3.1-3.15　研究方案确定

2015 3.16-4.15 模型的建立与算法实现

2015 4.16-5.15 实验与分析

2015 5.16-5.30 论文撰写

起讫日期 2015 年 3 月 1 日 至 2015 年 5 月 30 日

指导教师（签名）＿＿＿＿＿＿职称＿＿＿＿＿＿

三、系或研究所审核意见:

负责人（签名）＿＿＿＿＿＿

年　月　日

毕 业 论 文（设计） 考 核

一、指导教师对毕业论文（设计）的评语：

情感分析是当前的研究热点，作者以该方向为研究目标，选题具有较高的应用价值。作者在对情感分析的国内外研究现状进行深入分析的基础上，提出了基于 SVM 的微博情感分析方法。通过在人工标注数据集和 NLP 语料集上开展的实验分析表明，该方法具有一定的先进性和较高的实用性。论文工作表明作者具有扎实的基础理论知识，以及一定的科研工作的能力。论文条理清楚，层次分明，达到了本科毕业论文的要求。

指导教师(签名）

年 月 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述 占（10%）	开题报告 占（20%）	外文翻译 占（10%）	毕业论文（设计）质量及答辩 占（60%）	总 评成绩
分值					

答辩小组负责人（签名）

年 月 日