单位代码：10293 密 级：

南京邮电大学

硕士学位论文

论文题目：基于朴素贝叶斯的文本分类算法研究

学 号 1016020720

姓 名 何伟

导 师 张昀

学科专业 电路与系统

研究方向 智能信息处理

申请学位类别 工学硕士

论文提交日期 2018.2.18

Text Classification Algorithm Research Based on Naive Bayes

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Engineering

By

Wei He

Supervisor: Prof. Yun Zhang

February 2018

南京邮电大学学位论文原创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南京邮电大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

本人学位论文及涉及相关资料若有不实，愿意承担一切相关的法律责任。

研究生学号：_____ 研究生签名：_____ 日期：_____

南京邮电大学学位论文使用授权声明

本人承诺所呈交的学位论文不涉及任何国家秘密，本人及导师为本论文的涉密责任并列第一责任人。

本人授权南京邮电大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档；允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索；可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。本文电子文档的内容和纸质论文的内容相一致。论文的公布（包括刊登）授权南京邮电大学研究生院办理。

非国家秘密类涉密学位论文在解密后适用本授权书。

研究生签名：_____ 导师签名：_____ 日期：_____

摘要

互联网技术的飞速发展使得人们进入了大数据时代，互联网作为当今获取信息的主要渠道，与人类的关系也越来越密切。然后互联网中的绝大部分信息都是以文本形式存在，从而寻找一种能够有效处理文本数据进而对文本数据进行准确分类的方法成为当今具有重要研究价值的领域。朴素贝叶斯算法作为机器学习算法中的经典算法之一，以其模型简单、分类速度快、分类效率高等优点，成为了文本分类算法的重要研究内容。

对于朴素贝叶斯文本分类系统而言，一方面由于传统朴素贝叶斯理论是在假设了所有特征相互独立的基础上成立的，即特征词与特征词之间是相互独立的，这一定程度上影响了分类器的性能，因此如果能够寻找一些方法来削弱或消除特征独立性假设就可以相应的提高分类器的性能。另一方面对于海量的数据，如果不进行特征提取，就会增加分类系统的负担，降低分类器的性能，所以本文分别从文本分类系统的三个方向进行处理，提出了基于 IGDC 特征加权的朴素贝叶斯文本分类算法(IGDCNB)，基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)，改进的自定义特征维度的快速相关性过滤（IFSC-FCBF）算法。

本文的主要贡献：

（1）研究并改进了朴素贝叶斯特征加权算法模型，提出了基于 IGDC 特征加权的朴素贝叶斯文本分类模型。该模型通过全新的方式计算特征在每个类别和每个文档中两个维度的信息增益，并通过线性归一化的方式结合了两个维度的信息，大大削弱了朴素贝叶斯的特征条件独立性假设。

（2）研究了朴素贝叶斯深度加权模型，针对朴素贝叶斯模型自身的缺陷，改进了朴素贝叶斯模型条件概率的训练方式，将 IGDC 应用于朴素贝叶斯的深度加权中，提出了基于 IGDC 深度加权的朴素贝叶斯文本分类模型，进一步削弱了其特征条件独立性假设。

（3）首次将快速相关性过滤算法（FCBF）应用于文本分类中，综述了 FCBF 算法的应用领域及其在文本分类中存在的缺陷，改进了特征相关性的计算方式，并优化了原始 FCBF 算法步骤，提出了改进的自定义特征维度的快速相关性过滤（IFSC-FCBF）的朴素贝叶斯文本分类算法，在保证特征维度相同时，能够更加快速的选择出更加优越的特征，并且消耗更少的时间。

关键词：朴素贝叶斯；文本分类；特征加权；深度加权；快速相关性过滤

Abstract

The rapid development of Internet technology has made people enter the era of big data. As the main channel to obtain information today, the Internet is becoming more and more closely related to human beings. Then most of the information exists in the Internet is text data, so finding a method that can process text data effectively and classify text data accurately has become an important research field today. As one of the classical algorithms in machine learning algorithms, Naive Bayesian algorithm has become an important research content of text classification algorithms because of its simple model, fast classification speed and high classification efficiency.

For the naive Bayesian text classification system, on the one hand, the traditional naive Bayesian theory is based on the assumption that all features are independent of each other, that is, the feature words are independent of each other. It affects the performance of the classifier to some extends, so if you can find some ways to weaken or eliminate the feature independence assumption, the performance of the classifier will be improved accordingly.On the other hand, for the massive data, if the feature extraction is not implemented, this will increase the burden of the classification system and reduce the performance of the classifier. Therefore, the paper chooses the three directions of the text classification system, and proposes feature weighting with IGDC for Naive Bayes text classification (IGDCNB), deep weighting with IGDC for naive Bayesian text classification algorithm (IGDC-DWNB), improved feature size customized fast correlation-based filter for Naive Bayes (IFSC-FCBF) text classification.

The main contributions of this article:

(1) We researched and improved the naive Bayesian feature weighting algorithm model, and proposed a feature weighting with IGDC for Naive Bayes text classification (IGDCNB). The model calculates the information gain of features in each category and each document in a new way, and combines the information of two dimensions by linear normalization, which greatly weakens the feature independence hypothesis of naive Bayes.

(2) We researched the deep feature weighting model for naive Bayes and modifed the training method of the conditional probability of naive Bayes. On the same time IGDC is applied to the deep feature weighting of naive Bayes and proposed deep weighting with IGDC for naive Bayesian text classification algorithm (IGDC-DWNB). Experimental results show that the model can further weaken its feature condition independence hypothesis.

(3) It is the first time to use the fast correlation-based filter (FCBF) for text classification. The application fields of FCBF algorithm and its defects in text classification has been sumed up. We improved the calculation method of feature correlation and optimized its algorithm steps and proposed improved feature size customized fast correlation-based filter for Naive Bayes (IFSC-FCBF) text classification. In the same feature dimensions, the superior features can be selected more quickly and consumed less time.

Key words: Naive Bayes; text classification; feature weighting; deep weighting; fast correlation-based filter

目录

专用术语注释表

符号说明：

$\sum(\cdot)$	求和
$\prod(\cdot)$	求积
$max(\cdot)$	求最大值
$ln(\cdot)$	以自然对数为底取对数
$log(\cdot)$	以 2 为底取对数

缩略词说明：

TC	Text Classification	文本分类
VSM	Vector Space Model	向量空间模型
NB	Naive Bayes	朴素贝叶斯
BNB	Bernoulli Naive Bayes	伯努利朴素贝叶斯
MNB	Multinomial Naive Bayes	多项式朴素贝叶斯
MAP	Maximum a Posteriori	最大后验概率
TFIDF	Term Frequency–Inverse Document Frequency	词频反文档频率
TFIDFNB	Term Frequency–Inverse Document Frequency for Naive Bayes	TFIDF 加权朴素贝叶斯
IG	Information Gain	信息增益
MRMR	Minimum-Redundancy Maximum-Relevancy	最小冗余最大相关性
DFS	Distinguishing Feature Selector	判别式特征选择器
FCBF	Fast Correlation-Based Filter	快速相关性过滤器
IFSC-FCBF	Improved Feature Size Customized Fast Correlation-Based Filter	改进的自定义特征维度的快速相关性过滤器
IGDC	Information Gain of Documents and Category	二维信息增益
IGDCNB	Information Gain of Documents and Category Naive Bayes	基于 IGDC 加权的朴素贝叶斯
IGDC-DWNB	Deep Weighting with IGDC for Naive Bayes	基于 IGDC 的深度加权的朴素贝叶斯
TFIDF*IG	Term Frequency–Inverse Document Frequency with Information Gain	TFIDF 信息增益加权算法
DFWNB	Deep Feature Weighting Naïve Bayes	深度特征加权朴素贝叶斯算法

OFWNB	Ordinary Feature Weighting Naïve Bayes	普通特征加权朴素贝叶斯算法

第一章 绪论

1.1 论文的研究背景与意义

文本数据挖掘中的文本分类技术是一项实用价值较高、应用领域广泛的技术之一[1]。随着互联网的普及与快速发展，文本信息逐渐遍布于手机短信、各类电子邮件、网络图书馆及新闻网站等场所，由于参与人数的爆炸式增长其信息量也呈现出指数级的增长趋势。根据2018年8月中国互联网络信息中心（CNNIC）在京发布第42次《中国互联网络发展状况统计报告》中的数据显示[2]，截至2018年6月，我国网民规模为8.02亿，上半年新增网民2968万人，较2017年末增加3.8%，互联网普及率达57.7%。随着互联网的普及，文本信息的来源渠道也变得多种多样，每个互联网的参与者在获得信息的同时也在不断的贡献自己的信息资源。

如今我们正处于蓬勃发展的网络时代，其信息来源渠道广泛、巨大的信息容量、查询速度以及传播速度、更新速度快等特点让文本信息达到了空前的规模。尽管文本形式的信息使得用户获得消息变得更加便利以及可以看到来自世界各地各色各样的新闻娱乐科技信息，但是随之而来也有一些负面效应：如用户想要从铺天盖地的各种信息中精准的获得真正对自己有价值的信息变得越来越难，这对于信息检索技术的要求也相应的变得越来越高。因此，怎样有效地处理、分析这些海量信息，从中快速、准确地发现所需信息，已经成为当前信息科技领域一项非常有意义的课题。由此过去对信息的手动分类已经不能满足当前的用户需求，基于此，自动文本分类技术在当今起着尤为重要的作用。

文本分类技术是自然语言处理[3,4]的一个重要研究领域，其涉及到数据挖掘，信息检索，机器学习等多个领域[5,6,7,8,9]。自动文本分类是用算法对文本所包含的内容进行自动的分析，从而识别出不同文本的类别，通过对大量的文本数据进行批量处理就能快速判定对应类别从而准确的将文本数据进行归类处理，一方面可以使用户更方便的查询和阅读相关资料以得到自己想要的信息，另一方面也可以使各个新闻网站能够自动地快速整理相关文本信息，所以针对文本分类的研究具有重大的现实意义。

1.2 文本分类的研究现状

文本分类（Text Classification，TC）是机器学习领域的一个重要研究方向，即在给定的

分类标准下，将没有类别标签的文本，按照其内容将其划分到预先设定的类别中[10]。文本分类基本过程如图 1.1 所示。

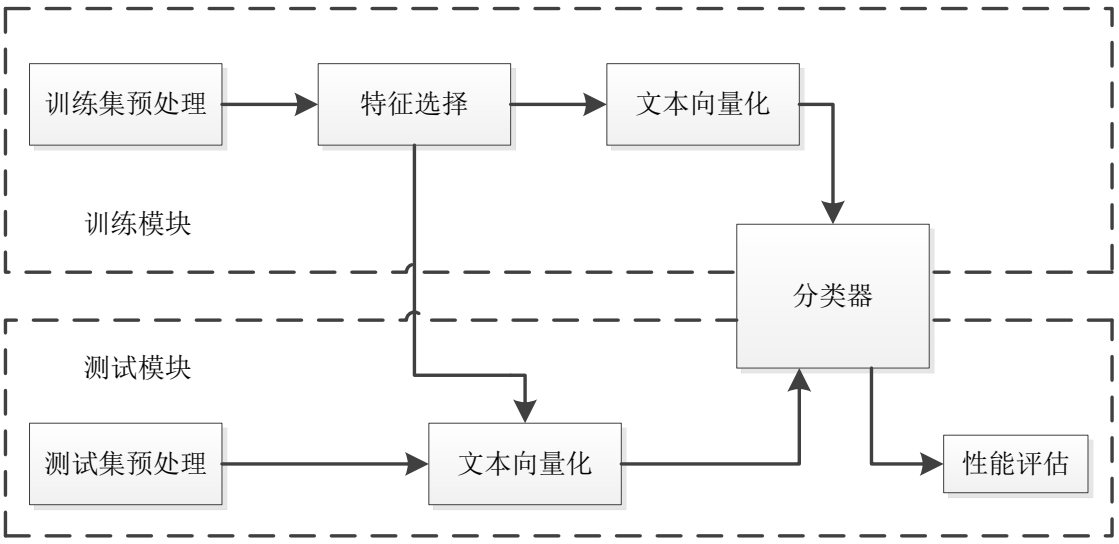

图 1.1 文本分类系统框图

最早开始研究文本分类公司是 IBM 公司，20 世纪 50 年代来自 IBM 的 H.P.Luhn 首次提出了统计词频的思想[11]，使用文本中的词频内容与文档建立联系的机制，能够初步的实现文本类别的判断，是文本分类起步的基础，对后来的文本分类起到了促进的作用。随后在 60 年代 Maron 和 Kuhns 共同首次提出了自动文本分类算法的思想[12]，这是自动文本分类的启蒙，对后来的文本分类起到了进一步的促进作用。到了 70 年代，Salton 在关于信息检索方面的论文中提出了向量空间模型[13]（Vector Space Model，VSM），之后向量空间模型在文本分类中得到了广泛的应用[14-18]。

进入 20 世纪 90 年代，为了能够更好地处理大量的电子文档，并且伴随着人工智能、机器学习、模式识别、统计理论等学科的发展，文本分类技术进入了自动分类时代。由于机器学习文本分类系统不需要先验知识且分类精度可以到达与专家相当的水平，同时机器学习方法的分类效率要显著优于领域专家，因此机器学习方法在文本分类领域得到了深入的研究和广泛的应用[19]，例如朴素贝叶斯[20-22]、K 近邻[23-25]、神经网络[26-28]、支持向量机[29-31]等。相比于这些方法，朴素贝叶斯有很多优势，主要包括其独特的不确定性知识表达形式、丰富的概率表达能力、综合先验知识的增量学习特性，其简洁性和有效性都要优于其他算法[19，32-35]。所以本文重点研究朴素贝叶斯算法。

朴素贝叶斯算法的思想是首先计算出各个类别的先验概率，再利用贝叶斯定理计算出各特征属于某个类别的后验概率，通过选出具有最大后验概率（maximum a posteriori, MAP）估计值的类别即为最终的类别，但朴素贝叶斯算法是在假设各个特征相互独立的条件下才成立的，这个假设显然不符合实际。为了削弱朴素贝叶斯特征条件独立假设的方法，学者们分别

从特征加权，算法改进和特征选择三个方向进行了改进[19]，提出了很多新方法。在特征加权方面，文献[36]中把特征信息增益加到 TFIDF 算法中相应改善算法性能后，之后文献[37]又把信息增益与信息熵结合，饶学者等人提出了根据特征在类间的词频和文档频率重新计算反文档频率[38]，吴学者等人把各个特征相对于类别的互信息作为权重进行加权处理[39]，蒋学者使用改进的信息增益率进行加权也取得良好的效果[40]，王学者等人利用特征选择方式，对选择出的特征给予固定的权重，对没有被选择的特征不给予权重的方式进行特征加权处理也使朴素贝叶斯算法得到了较大的提升[41]；国外学者也有采用加权方式提高朴素贝叶斯性能的算法[42-44]。然而这些算法的改进都比较单一，没有全面考虑影响特征权重的因素。本文从信息增益入手，考察特征词关于文档的信息增益和关于类别的信息增益，有效的结合了特征在两个方面的性能来刻画特征类别和特征文档对分类作用的提升程度，提出了基于特征二维信息增益的加权算法，对比其他算法有了明显的提升。

在算法改进方面，学者们大多使用深度加权的方式对朴素贝叶斯模型进行改进，如学者蒋首次提出了深度加权的方式来改进朴素贝叶斯算法[45]，之后另一位学者蒋再次使用 TFIDF 权重改进了深度加权的方式，与原始的普通加权方式相比，深度加权方式的提升效果较为明显[46]。所以本文再次从深度加权方式入手对朴素贝叶斯模型进行改进，提出了一种新的深度加权方式，相较于之前提出的普通加权方式，本文提出的深度加权方式又有了较大的提升。

在特征提取方面，机器学习中的主要有两种特征选择方法[47]：过滤器和包装器。过滤方法选择一个特征子集作为预处理步骤，它独立于分类算法工作。相反，包装方法需要分类器的精确度作为依据来进行特征选择。根据文献[48]所说，包装器往往能够得到更好的效果，因为对于预定义的算法它能够更好地选择特征子集。但是包装法具有更高的复杂度在选择特征时也需要更多的时间，对于文本分类任务显然是不可取的[49]。因此我们把关注点聚焦在过滤方法上。研究者们提出了许多特征过滤方法用于文本分类任务中，值得关注的包括文档频率[50]，信息增益[51]。然而与其他算法相比文档频率特征选择并不能取得好的效果,并且虽然信息增益能够很好的进行特征选择，它也有一个缺点，IG 只是根据特定的 IG 值进行筛选没有考虑特征之间的冗余[52]。为了能够有效的消除特征间的冗余，Peng 等人提出了消除冗余性的 MRMR[53],因其巨大的时间复杂度很难将其应用于文本分类中[52]。Lee 等人提出了改进信息增益特征选择算法[54]，Uysal 等人提出了基于特征概率性的选择方法：区别性特征选择算法（DFS）[55]。虽然这些算法能够比较有效的去除冗余，但都具有很高的复杂度，不能快速的进行特征选择。为了能更加快速的进行特征提取，我们重点研究了快速相关性过滤算法 FCBF[56]。FCBF 算法已经被应用在各个领域，YuLiu 等人将 FCBF 算法应用于采油机故障诊断上[57]，Chen 等人将 FCBF 和 Relief 结合用于基因选择[58]，Davood 等人将 FCBF

特征选择算法用于视频表情识别[59]，一些学者将 FCBF 算法用于医学方面，如 Sarojini 等人在糖尿病检测中用 FCBF 进行特征选择[60]，以及 BahaSEN 将其用于癫痫病诊断中的特征选择[61]。当前的研究成果表明，很少有特征选择方法在显著提高其分类精度的同时，还能维持它的模型简洁性和低时间复杂度，由于 FCBF 算法能够快速并有效的去除冗余特征，所以后续章节本文会重点研究 FCBF 算法。

1.3 本文的主要工作

本文的研究工作获得了国家自然科学基金“基于深度学习的移位 MIMO”鬼”成像方法”（项目批准号：61871234）的支持。本文的主要工作在于寻找更加有效的方法来削弱朴素贝叶斯的条件独立性假设，使朴素贝叶斯文本分类算法具有更好的鲁棒性。主要从特征加权，算法模型改进以及特征选择三个方面入手，具体的研究工作如下：

（1）从特征加权方面入手，本文第二章首先研究了朴素贝叶斯中常用的加权算法，并针对更加有效的信息增益(IG)加权算法的缺点进行改进，将特征与类别的信息增益，特征与文档的信息增益相结合，使用改进的方法计算特征对应的二维信息增益，给出了线性归一化的计算方式，提出了基于特征二维信息增益加权的朴素贝叶斯文本分类算法(IGDCNB)，与传统的 TFIDF 加权朴素贝叶斯算法以及 TFIDF*IG 算法相比，文本提出的算法在宏 F1 上都具有明显的优势。

（2）为进一步提升算法性能，本文第三章首先研究了深度加权的思想和优势，对前人[45,46]提出的深度加权方式进行改进，并使用第二章提出的 IGDC 算法计算每个特征词对应的信息增益进而进行深度加权，给出了改进的深度加权的计算公式，提出了基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)。与第二章提出的的 IGDCNB 算法以及文献[46]中提出的 DFWNB 算法和 OFWNB 算法相比，本文提出的深度加权算法在中文数据集和英文数据集上都具有更好的效果。

（3）本文前面的工作使用的特征选择算法都是最常用的文档频率特征选择算法，为了进一步提高朴素贝叶斯文本分类算法的性能，由于传统的文本特征选择算法都不能去除冗余特征，本文第四章从快速相关性过滤算法(FCBF)入手，验证了原始 FCBF 算法的弊端，改进了 FCBF 算法的相关性计算公式，并更新了算法的执行过程，提出了自定义特征维度的快速相关性过滤算法(IFSC-FCBF)。实验结果显示 IFSC-FCBF 算法与其他有效的文本特征选择算法 DFS[55]和 IG[51]比较时都具有更好的效果，同时具有最短的运行时间。

1.4 本文结构

本文针对图 1.2 的文本分类系统进行研究，图中的标号代表本文的三个主要创新点。

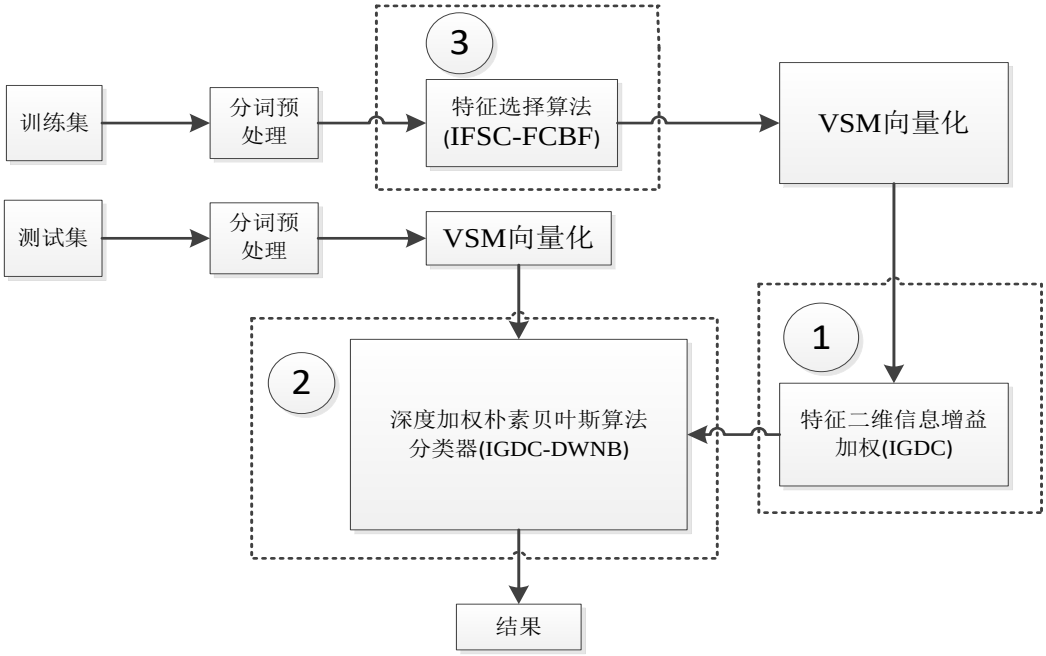

图 1.2 朴素贝叶斯文本分类系统框图

本文总共由五个章节组成，这五章的结构如下：

第一章是介绍背景知识的绪论部分，本章先介绍了本课题的研究背景，总结了朴素贝叶斯文本分类在国内和国外的发展进程以及改进的方向，最后总结本文的主要工作。

第二章介绍了朴素贝叶斯算法理论，文本分类过程中的预处理及算法性能验证方式，总结了针对朴素贝叶斯算法的特征条件独立性假设所做的常用的改进特征加权的算法，并结合信息增益加权的优点提出了基于特征二维信息增益加权的朴素贝叶斯文本分类算法(IGDCNB)。最后的实验结果显示，改进的基于二维特征信息增益加权算法与传统加权算法TFIDF[62]、TFIDF*IG[36]相比，在中文和英文文本分类上都有明显的提升。

第三章是介绍了常用的改进朴素贝叶斯模型的算法，以及传统的深度加权方式，给出了改进的深度加权方法的计算公式，并结合第二章提出的特征二维信息增益(IGDC)提出了基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)，实现结果显示与传统深度加权方式 OFWNB[55]以及文献[55]中的方法 DFWNB 相比，分类器的性能都有显著的提升。

第四章介绍了传统的快速相关性过滤算法(FCBF)，以及文本分类领域一些效果显著的特征提取算法，实验验证了原始的 FCBF 算法并不适用于文本分类，所以本章对其相关性计算

方法进行改进，给出了新的计算方式，并优化算法的执行步骤，提出了自定义特征维度的快速相关性过滤算法(IFSC-FCBF)，实验结果显示本文提出的特征选择算法不仅能够提升分类器的性能，同时具有更低的时间复杂度。

第五章是最后的总结部分。总结全篇内容，并对本课题未来的研究重点和有待完善的部分做了展望。

第二章 基于特征二维信息增益加权的朴素贝叶斯文本分类算法

朴素贝叶斯分类模型由于其简单性、高效性和有效性被广泛由于解决文本分类问题[19]，其算法思想是首先计算出各个类别的先验概率，再利用贝叶斯定理计算出各特征属于某个类别的后验概率，通过选出具有最大后验概率（maximum a posteriori, MAP）估计值的类别即为最终的类别。朴素贝叶斯文本分类问题需要考虑特征之间的独立性，属于离散型问题，离散型朴素贝叶斯模型分为基于二项分布的伯努利朴素贝叶斯以及基于多项式分布的多项式朴素贝叶斯。由于 伯努利朴素贝叶斯模型在统计一篇文档时只是简单地考虑单词出现与否，而没有考虑单词出现的频率，因此该模型的分类精度会受到影响。先前的研究[19,21,41]表明多项式朴素贝叶斯模型在大数据集上往往表现得更好，并且总体而言多项式朴素贝叶斯模型在分类精度上要优于伯努利朴素贝叶斯模型，所以本文采用的朴素贝叶斯模型为多项式朴素贝叶斯模型。

本章首先详细阐述了朴素贝叶斯两种模型的区别，以及文本语料库的预处理、算法性能的评估方法，接着介绍了常用的加权方法，结合文献[52,63]中提到的信息增益加权是比较优越的加权方式及思想，提出了 IGDCNB(Information Gain of Documents and Category Naive Bayes)算法，给出了信息增益的计算方式。在本章的最后通过仿真实验验证了新提出的 IGDCNB 算法相较于传统的 TFIDF*IGNB 以及 TFIDFNB 在特征维度相同时具有更高的 F1 值与宏 F1 值。

2.1 朴素贝叶斯算法

2.1.1 伯努利朴素贝叶斯算法模型

伯努利模型[64]（Bernoulli Naïve Bayes，简称 BNB）认为一个事件有两种可能性，发生或者不发生。当进行 n 次独立重复的伯努利试验，会产生一个新的分布称为二项分布。对于一篇文档而言，词典中的每个单词可能在文档中出现、也可能不在文档中出现，因此对于词典中的一个单词可以看作是进行一次伯努利试验，而词典中的所有单词可以看作 n 重伯努利试验，就是二项分布。对于一篇文档 d，表示为向量形式 $d=\{t_1,t_2,t_3...t_m\}$，$t_k\in\{1,0\}$，其中

$t_k=1$表示该单词在文档 d 中出现，反之则未出现，m 表示词典的大小。为了处理文本数据，朴素贝叶斯的一个主要假设是在给定文档类别的情况下，每个单词条件概率计算是相互独立的，在此假设下，BNB 模型可以使用公式 2.1 来预测文档 d 的类：

$$c(d)=\arg\max_{c\in C} p(C_j)\prod_{i=1}^{m}(t_k p(Dt_k\mid C_j)+(1-t_k)(1-p(Dt_k\mid C_j))) \quad (2.1)$$

其中 $p(C_j)$ 表示先验概率， $p(Dt_k\mid C_j)$ 表示条件概率，可以采用频数计数近似估计，计算公式分别如下所示：

$$p(C_j)=\frac{tf(D,C_j)}{\sum_{j=1}^{V} tf(D,C_j)} \quad (2.2)$$

$$p(Dt_k\mid C_j)=\frac{tf(Dt_k,C_j)}{tf(D,C_j)} \quad (2.3)$$

其中 $tf(Dt_k,C_j)$ 表示含有单词 t_k 的文档在 C_j 类中出现的文档数， $tf(D,C_j)$ 为 C_j 类中的所有文档数，为了避免概率计算时出现 $p(Dt_k\mid C_j)$ 为 0 的情况，一般采用拉普拉斯平滑处理，具体计算公式如下：

$$P(Dt_k\mid C_j)=\frac{\sum_{i=1}^{n} t_{ki}\delta(C_j,C)+1}{\sum_{i=1}^{n}\delta(C_j,C)+2} \quad (2.4)$$

其中 $\delta(\bullet)$ 表示二值函数其公式如下：

$$\delta(x,y)=\begin{cases}1, x=y\\ 0, x\neq y\end{cases} \quad (2.5)$$

2.1.2 多项式朴素贝叶斯模型

多项式模型[65]（Multinomial Naïve Bayes 简称 MNB）将文档看作一个词袋模型，认为单词在一篇文档中出现的频率对文档类别的预测有影响。因此在计算条件概率的时候，MNB 需要统计单词出现的频率，这一点与 BNB 有显著的不同。设文档类别为 $C=\{C_1,C_2...C_j\}$，$j=1,2,3...V$，设 D_i 为任意一篇文档，其包含的 m 个特征词为 $D_i=\{t_1,t_2....t_m\}$，其对应的最大的后验概率的类别即为文档 D_i 的所属的类别，后验概率公式可表示为：

$$P(C_j\mid D_i)=\frac{P(D_i\mid C_j)P(C_j)}{P(D_i)} \quad (2.6)$$

其中，$P(C_j)$表示类别C_j出现的概率，$P(D_i|C_j)$表示文档D_i属于类别C_j的条件概率，$P(D_i)=P(t_1,t_2...t_m)$表示所有特征的联合概率。

贝叶斯分类的过程就是求解$P(C_j|D_i)$最大值的过程，显然对于给定的训练文档$P(D_i)$是个常数。所以求解过程可转化成求解式 2.7。

$$C_{map}=\max_{C_j\in C}P(C_j|D_i)=\max_{C_j\in C}P(C_j)P(D_i|C_j) \qquad (2.7)$$

其中C_{map}表示最终的分类结果。

根据朴素贝叶斯的条件独立性假设，所以（2）式可简化为式 2.8。

$$C_{map}=\max_{C_j\in C}P(C_j|D_i)=\max_{C_j\in C}P(C_j)P(\{t_1,t_2...t_m\}|C_j)$$

$$=\max_{C_j\in C}P(C_j)\prod_{k=1}^{m}P(t_k|C_j) \qquad (2.8)$$

普通加权朴素贝叶斯模型为式 2.9 所示。

$$C_{map}=\max_{Cj\in C}P(C_j|D_i)=\max_{Cj\in C}P(C_j)\prod_{k=1}^{m}P(t_k|C_j)^{W_k} \qquad (2.9)$$

由于每次计算的概率可能会比较小，为了避免出现下溢的情况，通常采用对决策规则取对数的形式，最终的判别方式为式 2.10。

$$C_{map}=\max_{Cj\in C}P(C_j|D_i)=\max_{Cj\in C}[lnP(C_j)+\sum_{k=1}^{m}lnP(t_k|C_j)\times W_k\times TF_{t_k}] \qquad (2.10)$$

式中，m表示特征单词数，t_k（$k=1,2,3....m$）表示文档D_i中的第k个特征词；先验概率$P(C_j)$和条件概率$P(t_k|C_j)$的计算公式如 2.11,2.12 所示。

$$P(C_j)=\frac{\sum_{i=1}^{n}\delta(C_i,C_j)+1}{n+V} \qquad (2.11)$$

$$P(t_k|C_j)=\frac{\sum_{i=1}^{n}TF_{it_k}\delta(C_i,C_j)+1}{\sum_{k=1}^{m}\sum_{i=1}^{n}TF_{it_k}\delta(C_i,C_j)+m} \qquad (2.12)$$

式中n表示总的训练文档数，V表示类别个数，C_i表示第i篇训练文档的类别，TF_{it_k}表示特征词t_k在文档D_i中的出现频数，$\delta(\bullet)$表示二值函数。

2.2 文本预处理及分类器性能评估

对于中文数据集，本文使用通用的 JIEBA 分词库进行中文分词[46]，对于英文数据集，可直接使用英文书写中的空格进行分割，在得到特征词库之后，再统一去除低频词语及停用

词。停用词是一些没有意义的虚词和标点符号，中文的包括如‘的’，‘啊’，‘比如’，‘不但’等之类的词，英文的包含如‘an’，‘about’，‘for’，‘if’等之类的词。停用词一般是根据领域专家构建的停用词表进行去除。在文本信息完成分词过后，需要使用向量空间模型(VSM)进行文本向量化。在向量空间模型中，一篇文档对应一个高维的向量，而文档中的每个单词对应向量中的一个属性。通常向量的维度为文档数据集中出现的所有单词的数目，而向量每一维的取值为该单词在文档中出现的频率。如“Everything will be ok in the end, if it's not ok, it's not the end”经过分词之后得到这样一个词库[be, ok, in, will, the end, everything, if, it’s, not]，接着统计该段文档中所有出现的词频，得到向量化过后的文档表示如下表 2.1 所示。

表 2.1　向量化后的文档表示

be	ok	in	will	the	end	everything	if	it’s	not
1	2	1	1	2	2	1	1	1	2

在分类器对文本完成分类后，需要对分类器性能进行评估。分类器的评估分为闭测试和开测试两种方式。闭测试就是将训练集的文本即使分类器的训练文本，同时还是训练集的测试文本。开放测试，则是测试集的文本与训练集的文本相互独立，没有交集。一般文本数据集较多时采用开放测试。常用的文本分类器性能评估参数有：查全率（Recall）、查准率（Precision）、F1 值、宏平均[40，45，46，52]等。

若有如下表 2.2 所示混淆矩阵，其中 TP 表示正确的标记为正，FP 错误的标记为正，FN 错误的标记为负，TN 正确的标记为负。

表 2.2 混淆矩阵

真实情况	预测情况	
	正例	反例
正例	TP	FN
反例	FP	TN

则查全率和召回率的计算公式为 2.13,1.24 所示。

$$P = \frac{TP}{TP + FP} \quad (2.13)$$

$$R = \frac{TP}{TP + FN} \quad (2.14)$$

从上式中可以看出查准率反应的是分类器将判别文本分为某类，而该文本确实属于本类的概率。查全率反应的是文本属于该类，并且分类器正确分类的概率。两个指标分别描述分类器的分类完整性和正确性。

查全率和查准率都是针对分类器在单个类别上的评估参数，只能反映分类器的局部性能，F1值是衡量查准率和召回率大小的综合指标，其计算公式如2.15所示。

$$F1=\frac{2*P*R}{P+R} \quad (2.15)$$

对于多分类问题，宏平均是对分类器的整体性能进行评估，宏查全率，宏召回率及宏F1值的计算公式如2.16，2.17，2.18所示

$$Macro\_P=\frac{1}{V}\sum_{i=1}^{V}P\text{，}V\text{表示类别数} \quad (2.16)$$

$$Macro\_R=\frac{1}{V}\sum_{i=1}^{V}R \quad (2.17)$$

$$Macro\_F1=\frac{2*Macro\_P*Macro\_R}{Macro\_P+Macro\_R} \quad (2.18)$$

2.3 朴素贝叶斯算法中的权重计算方法

由于朴素贝叶斯的特征条件独立性假设影响了分类器的性能，所以学者们大多数使用特征加权的方式来抑制这个假设。特征词权重计算的核心思想就是对根据特征词选择算法选择出来的特征词利用某种权重计算方法赋予不同的权重，对分类贡献能力大的特征词赋以较高权重，而对分类能力较差的特征词赋予较低权重，从而让特征词具有更好的区分文本类别的能力，以达到提高分类的准确率的目的。

2.3.1 词频法

词频（Term Frequency，TF），即特征词在文档中出现的频数，权重的计算即为特征词在文档中出现的频数，所以对于特征词t_k，其在文档D_i中的权重为式2.19。

$$w_{t_k}=TF(t_k,D_i) \quad (2.19)$$

在中英文文本中，由于存在大量的虚词，也称为停用词，在用词频赋予权重时，往往是这些没有意义的停用词具有较大的权重，所以词频法的缺点很明显。

2.3.2 反文档频率法

反文档频率法（Inverse Document Frequency，IDF）也称为逆文档频率，其思想为如果特征词在某一类文档中出现较多，则其对于该类文档具有更高的分类能力。反文档频率的计算

公式如 2.20 所示。

$$IDF(t_k) = log(\frac{N}{n(t_k)} + 0.01) \quad (2.20)$$

其中 $IDF(t_k)$ 表示特征词 t_k 的反文档频率，N 表示总文档数，$n(t_k)$ 表示含有特征词 t_k 的文档数目。反文档频率方法的缺点是该方法是从整体角度来衡量特征词的重要程度，而没有考虑特征词在类内的分部信息

2.3.3 TFIDF 权重法

TFIDF 权重是在文本分类中运用比较多的计算方法，是由 Salton 提出[62]。TFIDF 算法的思想：特征单词在某特定文本中出现的频数越大，其对于该文本的分类作用越大，特征单词在大多数文档中出现的频数越大，对于文本的分类作用越小，其结合了词频与反文档频率两者的优点。TFIDF 算法将词频和反文档频率结合作为特征的权重，归一化计算方法如 2.21, 2.22 所示。

$$IDF(t_k) = lb(\frac{N}{n(t_k)} + 0.01) \quad (2.21)$$

$$W_k = TF(t_k) * IDF(t_k) = \frac{TF(t_k) * IDF(t_k)}{\sqrt{\sum_{i=1}^{m}(TF(t_k) * IDF(t_k))^2}} \quad (2.22)$$

其中 $TF(t_k)$ 为特征 t_k 在训练集中出现的频数，$IDF(t_k)$ 是反文档频率，N 表示训练集的总文档数，$n(t_k)$ 表示出现特征 t_k 的文档数。同样的 TFIDF 算法考虑了特征词的局部和全局的分布特性，但并没有考虑特征词在类内和类间的分布情况【32-33】。

2.3.4 信息增益权重法

信息增益是描述特征在得知某一属性之后的信息熵的变化量，在文本分类中表现为特征词在确定了所属类别之后的信息熵的变化量。其计算公式如 2.23 所示。

$$\begin{aligned} IG(t) &= H(C) - H(C|t) \\ &= -\sum_{j=1}^{V} P(C_j)\log(P(C_j)) + P(t)\sum_{j=1}^{V} P(C_j|t)\log(P(C_j|t)) \\ &\quad + P(\bar{t})\sum_{j=1}^{V} P(C_j|\bar{t})\log(P(C_j|\bar{t})) \end{aligned} \quad (2.23)$$

其中 $IG(t)$ 表示特征词 t 的信息增益值，C_j 表示文本类别，$H(C)$ 表示的是在没有得到特征词 t 时文本的类别信息熵。 $H(C|t)$ 为获得特征词 t 后文本属于某个类别的信息熵。这种在得到特征词 t 前后，文档的类别信息熵的减少量就是信息增益，它蕴含的是特征词 t 对分类所能提供信息量的大小。

将信息增益的概念应用到文本分类领域，其基本原理如下：假设文本集合是符合某种分布规律的数据源，计算系统的信息熵与文本中特征词的条件熵之间的差值，即信息增益来确定特征词所携带的信息量，信息增益值越大则该特征词携带的分类信息越多，在分类过程中越重要，在文本表示时应该赋予较高的权重。反之则该特征词携带的信息量较小，文本表示时应该赋予较低权重。

2.3.5 TFIDF*IG 的加权算法

针对 TFIDF 的缺陷，张学者[66]把信息增益引入到 TFIDF 算法中提出了 TFIDF*IG 算法，首先计算出各个类别的信息熵，然后计算各特征词在每个类别中的条件信息熵，利用两者的差值计算出单词在各个类别中的信息增益，把该信息增益反应在权重中，计算公式如 2.24,2.25 所示。

$$W_i = TF(t_i) * IDF(t_i) * IG(t_i, C_j) \qquad (2.24)$$

$$IG(C,t) = E(C) - E(C_j|t) = \sum_{j=1}^{V} P(C_j|t) * lb(P(C_j|t)) - \sum_{j=1}^{V} P(C_j) * lb(P(C_j)) \qquad (2.25)$$

其中 C 为文档的类别集合，$P(C_j)$ 为类别 C_j 在训练集中的概率，$P(C_j|t)$ 为每个特征词 t 在类别 C_j 中出现的概率，V 表示类别总数。

利用 TFIDF*IG 算法能够将特征在类别中的信息反应出来，并同时能够对每个特征权重做一定的修正。当特征词 t 在某个类别中分布很多，而在其他类别中分布很少时，利用信息增益计算公式就能得到很高的信息增益值，这样就能很好的反应出特征词的分布对分类的影响，反之就能得到较小的信息增益值，所以在一定程度提高了算法的精确度。

2.4 特征二维信息增益(IGDC)加权朴素贝叶斯算法

在 2.3 中对传统的特征词权重计算方法进行了深入的研究与分析，通过算法分析可知，传统的 TFIDF 算法忽略了特征词的类内、类间分布信息对文本分类结果的影响[66]。因此，

为了克服该算法的不足之处，提高文本分类的准确性，本章提出了一种特征二维信息增益(IGDC)加权朴素贝叶斯算法。该算法在保证相同的特征维度时，能具有更好的效果。

由于 2.3.5 给出的 TFIDF*IG 算法也只考虑了特征词在类列间的分布情况并没有考虑到特征词在每个类别文档中的出现情况，因此会对对权重造成的影响。以进一步提高算法精度为目标，针对 TFIDF*IG 算法的缺陷，这里定义一个新的权重计算函数：IGDC 函数。由于信息增益是描述某个属性对分类效果提升作用的指标，信息增益越大，意味着特征属性对文档分类提升越大[67]。所以本章从特征二维信息增益入手，考察特征词关于文档的信息增益和关于类别的信息增益，有效的结合了特征在两个方面的性能来刻画特征类别和特征文档对分类作用的提升程度，式 2.26 定义了新的方法求特征类别概率。

$$P(t,C_j)=\frac{tf(t,C_j)+L}{\sum_{j=1}^{V}tf(t,C_j)+V*L} \quad (2.26)$$

$tf(t,C_j)$表示各特征词在C_j类中的频数，所以$P(t,C_j)$就表示C_j类中出现的特征词在训练集该特征词总数中出现的概率。式中的L是为了抑制概率为 0 的情况所加入的平滑因子，本文中取L=0.01，V表示类别数。同样的方法，式 2.27 得到各类别中特征词出现的文档数在训练集中对应特征词所出现的总文档数中出现的概率。

$$P(D_t,C_j)=\frac{tf(D_t,C_j)+L}{\sum_{j=1}^{V}tf(D_t,C_j)+V*L} \quad (2.27)$$

$tf(D_t,C_j)$表示在C_j类中含有特征词t的文档数，L=0.01 为平滑因子，V表示类别数。

传统的求特征文档信息增益的方法仅仅考虑了特征与文档的关系[15]，而忽略了文档与文档类别的关系，所以这里定义新的求特征文档信息增益的公式把特征与文档的关系同时把文档与文档类别的关系结合在一起，由此可以得到新的特征类别信息增益和特征文档信息增益如 2.28 和 2.29 所示。

$$IGC=E(C_j)-E(C_j\mid t)=$$
$$\sum_{j=1}^{V}P(t,C_j)\log P(t,C_j)-\sum_{j=1}^{V}P(C_j)\log P(C_j) \quad (2.28)$$

$$IGD=E(C_j)-E(C_j\mid D_t)=$$
$$\sum_{j=1}^{V}P(D_t,C_j)\log P(D_t,C_j)-\sum_{j=1}^{V}P(C_j)\log P(C_j) \quad (2.29)$$

其中，IGC 表示特征类别信息增益，刻画特征与类别的关系；IGD 表示特征文档信息增益，刻画特征与文档的关系；$P(t,C_j)$和$P(D_t,C_j)$分别表示上文提出的求特征类别概率和特征文档概率。这样得到的两组信息增益能够准确的反应出每个特征词对每个类别的影响力以及每个特

征词对每类文档的影响力。同时把特征词类别信息增益和文档信息增益结合起来，并采用线性归一化方法进行处理，得到权重表示为式 2.30。

$$W_{IGDC}=\frac{IGD*IDC-\min(IGD*IGC)}{\max(IGD*IDC)-\min(IGD*IGC)} \qquad (2.30)$$

W_{IGDC} 便是组合了两种信息增益后得到的特征二维信息增益，线性归一化是为了使数据等比例的变化，这样不会影响整体的权重调整。

下面举例说明新权重的合理性，假设训练集包含 3 个类别，每个类别中有 3 篇文档，特征词集合为$\{t_1,t_2,t_3\}$，分布情况如下表 2.3 所示。

表 2.3 特征词分布

	类别 1			类别 2			类别 3		
特征词	1	2	3	1	2	3	1	2	3
t_1	5	5	5	0	0	0	0	0	0
t_2	8	0	0	0	8	0	0	0	8
t_3	0	0	0	0	6	0	0	10	0

由表 1 知t_1只在类别 1 中出现过，且在三个文本中都出现过，说明t_1能够正确的代表类别 1 的信息，应当给予较大的权重，t_2在三个类别中都出现相同的次数，说明不具有分类能力，应当给予较小的权重，t_3大部分出现在类别 3 中，所以分类能力要比t_2好，但比t_1要差，所以权重值应当介于t_1和t_2之间，使用以上三个算法得到的权重结果如下表 2.2 所示。

表 2.4 特征加权算法结果比较

加权算法	各特征单词权重		
	t_1	t_2	t_3
TFIDF	0.419	0.671	0.612
TFIDF*IG	0.865	0.000	0.502
IGDC	1.000	0.000	0.142

由表 2 中的结果可以看出，TFIDF 算法因为针对的是整个训练集中的特征，所以词频越大的特征被分配的权重越大，导致结果与实际情况有点截然相反。TFIDF*IG 算法考虑了特征词与类别间的关系，所以权重分配比较合理，但因为仍然与反文档频率相结合导致t_1与t_3的权重相差很小，这种时候可能会影响到分类效果，相比之下 IGDC 算法不仅让没有分类能力的特征词t_2权重消零，而且让t_1与t_3的权重拉开了差距，这样能让各个特征起到决定分类作用的效果。

2.5 实验设计与结果分析

本节是为了设计实验来验证本章中提出的算法的分类性能，分别在四个数据集上，两个英文数据集：20newsgroup[68,69], Ruster21578[68，70]和两个中文数据集：搜狗实验室语料库[71]和复旦大学中文语料库[46]。20newsgroup 数据集收集了大约 20,000 左右的新闻组文档，均匀分为 20 个不同主题的新闻组集合，Ruster21578 包含 36 个类别，一共 10835 个文档；搜狗实验室语料库是从 sohu 新闻整理得到，含有 10 个新闻类别，每个类别含有 2000 个文档，复旦大学中文语料库是由复旦大学的李荣陆提供同样含有 10 个类别含有 2816 个文档。由于 Ruster21578 的文档分布极其不均匀，所以从中选出 6 个类别，每个类别选出 150 文档，一共 900 文档，在其他语料库中选出 6 个类别，每个类别选出 200 文档，一共 1200 文档；对比了传统的 TFIDF 算法[62]，TFIDF*IG 算法[66]。其中数据集分布如下表 2.5 所示。

表 2.5 数据集分布

数据集	文档数	训练集	测试集	类别数
20_newsgroup	1200	720	480	6
Ruster21578	900	540	360	6
搜狗实验室语料库	1200	720	480	6
复旦大学语料库	1200	720	480	6

在本章的实验中，分别设计四组实验分别在 20newsgroup，Rusters21578，搜狗实验室语料库及复旦大学语料库四个不同的数据集上进行验证，比较本章中介绍的三个算法：基于改进特征二维信息增益加权的朴素贝叶斯文本分类算法(IGDCNB)，基于信息增益的 TFIDF 加权的朴素贝叶斯文本分类算法(TFIDFIGNB)，基于 TFIDF 加权的朴素贝叶斯文办分类上算法(TFIDFNB)。实验语言使用 Python3.6，在处理器为 i5-4210，频率为 2.60GHz，内存为 8G，操作系统为 win10 的笔记本上进行。以文档频率作为选取特征词的标准每次实验进行 10 次交叉验证取平均值，在四个数据集上的查准率 P，召回率 R 以及 F1 值如下表 2.6-2.9 所示。

表 2.6 20newsgroup 上的结果，特征维度为 500

	基于特征二维信息增益 IGDC 加权的朴素贝叶斯文本分类算法 (IGDCNB)			**基于 TFIDF*IG 特征加权的朴素贝叶斯文本分类算法(TFIDF*IGNB)**			**基于传统的 TFIDF 特征加权的朴素贝叶斯文本分类 (TFIDFNB)**		
类别	P	R	F1	P	R	F1	P	R	F1
alt.atheism	0.9524	0.9302	**0.9412**	0.9277	0.8953	0.9112	0.9610	0.8605	0.9080
comp.graphics	0.9041	0.7674	0.8302	0.8706	0.8605	**0.8655**	0.8961	0.8023	0.8466
misc.forsale	0.8571	0.8571	0.8571	0.8732	0.8052	0.8378	0.9155	0.8442	**0.8784**
rec.autos	0.8462	0.7432	**0.7914**	0.7564	0.7973	0.7763	0.7108	0.7973	0.7516

sci.crypt	0.8421	0.9195	**0.8791**	0.8929	0.8621	0.8772	0.9241	0.8391	**0.8795**
talk.politics.guns	0.7791	0.9571	**0.8590**	0.7975	0.9000	0.8456	0.7097	0.9429	0.8098
平均值	**0.8635**	**0.8624**	**0.8600**	0.8530	0.8534	0.8528	0.8528	0.8477	0.8456

表 2.7　Rusters21578 数据集结果，特征维度 500

	基于特征二维信息增益 IGDC 加权的朴素贝叶斯文本分类算法 (IGDCNB)			**基于 TFIDF*IG 特征加权的朴素贝叶斯文本分类算法(TFIDF*IGNB)**			**基于传统的 TFIDF 特征加权的朴素贝叶斯文本分类 (TFIDFNB)**		
类别	P	R	F1	P	R	F1	P	R	F1
Crude	0.9487	0.9610	**0.9548**	0.9848	0.8442	0.9091	0.9577	0.8831	0.9189
Acq	0.9000	0.9000	**0.9000**	0.8515	0.9556	**0.9005**	0.8810	0.8222	0.8506
Interest	0.6422	0.8537	**0.7330**	0.5868	0.8659	0.6995	0.5833	0.8537	0.6931
Wheat	0.9643	0.9878	0.9759	0.9759	0.9878	0.9818	0.9878	0.9878	**0.9878**
Money	0.7451	0.5135	**0.6080**	0.6957	0.4324	0.5333	0.6600	0.4459	0.5323
Earn	0.9412	0.8533	**0.8951**	0.9524	0.8000	0.8696	0.8356	0.8133	0.8243
平均值	**0.8569**	**0.8448**	**0.8444**	0.8411	0.8143	0.8156	0.8175	0.8010	0.8011

表 2.8　搜狗数据集结果，特征维度 500

	基于特征二维信息增益 IGDC 加权的朴素贝叶斯文本分类算法 (IGDCNB)			**基于 TFIDF*IG 特征加权的朴素贝叶斯文本分类算法(TFIDF*IGNB)**			**基于传统的 TFIDF 特征加权的朴素贝叶斯文本分类 (TFIDFNB)**		
类别	P	R	F1	P	R	F1	P	R	F1
健康	0.9118	0.7470	**0.8212**	0.8507	0.6867	0.7600	0.8500	0.6145	0.7133
教育	0.7609	0.9333	**0.8383**	0.8485	0.7467	0.7943	0.8261	0.7600	0.7917
军事	0.8913	0.9213	**0.9061**	0.8119	0.9213	0.8632	0.8125	0.8764	0.8432
旅游	0.7849	0.9733	**0.8690**	0.7609	0.9333	0.8383	0.7216	0.9333	0.8140
体育	0.9841	0.7654	**0.8611**	0.9130	0.7778	0.8400	0.9000	0.7778	0.8344
文化	0.7917	0.7403	**0.7651**	0.6941	0.7662	0.7284	0.6818	0.7792	0.7273
平均值	**0.8541**	**0.8467**	**0.8434**	0.8131	0.8053	0.8040	0.7986	0.7902	0.7873

表 2.9 复旦大学语料库，特征维度 500

	基于特征二维信息增益 IGDC 加权的朴素贝叶斯文本分类算法 (IGDCNB)			**基于 TFIDF*IG 特征加权的朴素贝叶斯文本分类算法(TFIDF*IGNB)**			**基于传统的 TFIDF 特征加权的朴素贝叶斯文本分类 (TFIDFNB)**		
类别	P	R	F1	P	R	F1	P	R	F1
环境	0.8974	0.8642	**0.8805**	0.9306	0.8272	0.8758	0.9701	0.8025	0.8784
交通	0.8488	0.9125	**0.8795**	0.8750	0.7875	0.8289	0.8971	0.7625	0.8243
教育	0.8367	0.9880	0.9061	0.8778	0.9518	**0.9133**	0.8778	0.9518	**0.9133**

军事	0.9242	0.7093	**0.8026**	0.8750	0.6512	0.7476	0.8966	0.6047	0.7222
经济	0.8846	0.8415	**0.8625**	0.8000	0.8293	0.8144	0.7907	0.8293	0.8095
体育	0.8514	0.9265	**0.8873**	0.6804	0.9706	0.8000	0.5946	0.9706	0.7374
平均值	**0.8738**	**0.8736**	**0.8697**	0.8398	0.8362	0.8300	0.8378	0.8202	0.8141

从表2.6和表2.7中可以看出本文提出的特征二维信息增益(IGDC)加权方式在两个英文数据集上在大多数类别上的查准率，召回率和F1值三个指标上都要高于其他两个加权算法，其中在20newsgroup数据集上平均F1值比TFIDFIG加权朴素贝叶斯算法和TFIDF加权朴素贝叶斯算法分别高出0.72%，1.44%，在Rusters21578数据集上比同样的高出2.88%，4.33%。特别的对于alt.atheism，Crude，Interest ，Money都高出了5个百分点左右。

从2.8和表2.9中可以看出在中文数据集上，加入了类间信息增益的TFIDFIG加权算法比传统的TFIDF加权算法在六个类别的F1值上都有不错的提升，而加入了类间信息增益和文档信息增益的IGDC加权算法在所有类别上都有大幅的提升。这就是因为IGDC结合的是特征与类别，特征与文档之间的信息增益，并采用了线性归一化处理的方式，保证了不相关的特征与相关的特征能够更加明显，从而能够更好的起到决定分类效果的作用。在搜狗语料库上IGDC加权算法的平均F1值要比传统的TFIDFIF和TFIDF加权算法要分别高出3.94%，5.61%，在复旦大学语料库上要分别高出3.97%和5.56%，充分显示了本文提出的算法在中文数据集以及英文数据集上的有效性。

为了研究特征维度与准确率之间的关系，我们在不同的特征维度下进行了实验，来观察三个算法的宏F1值的在不同数据集上的变化，如下图2.1-2.4所示。

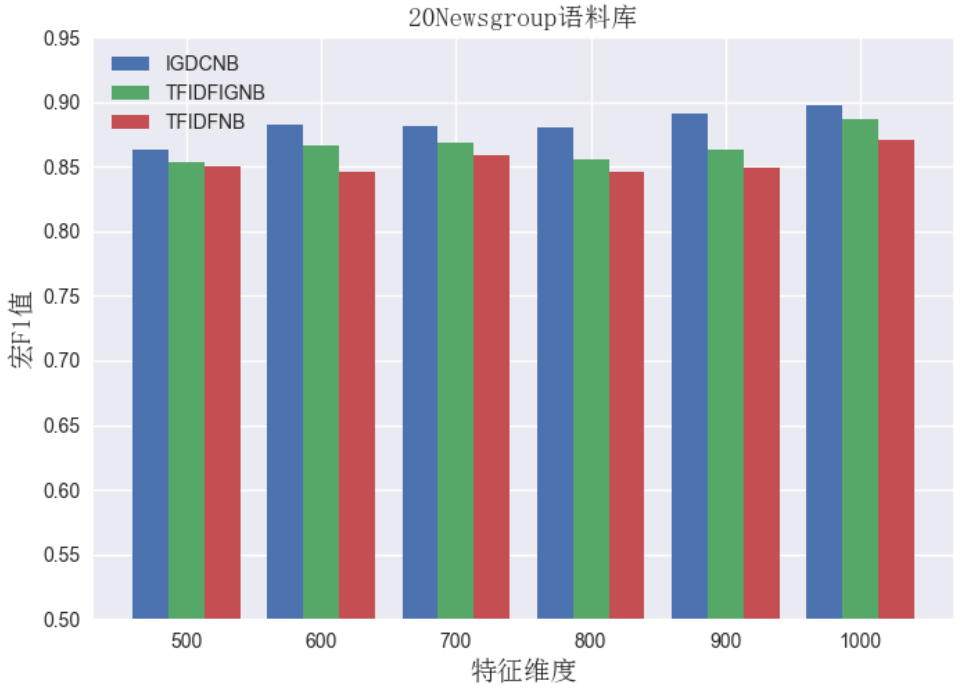

图 2. 1 20newsgroup 语料库各算法宏 F1 值对比

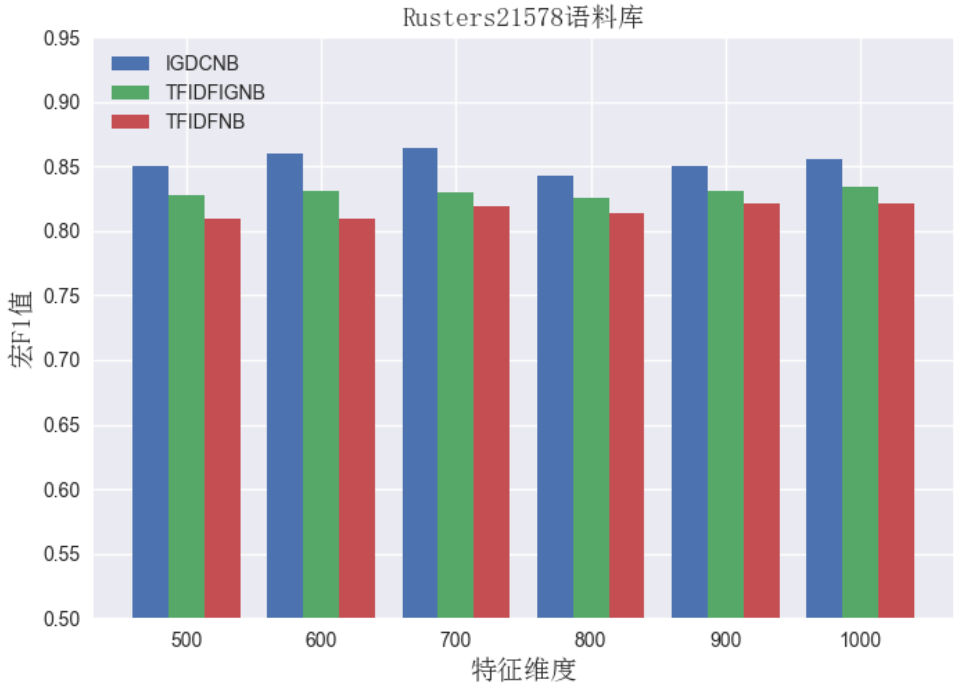

图 2.2 Rusters21578 语料库各算法宏 F1 值对比

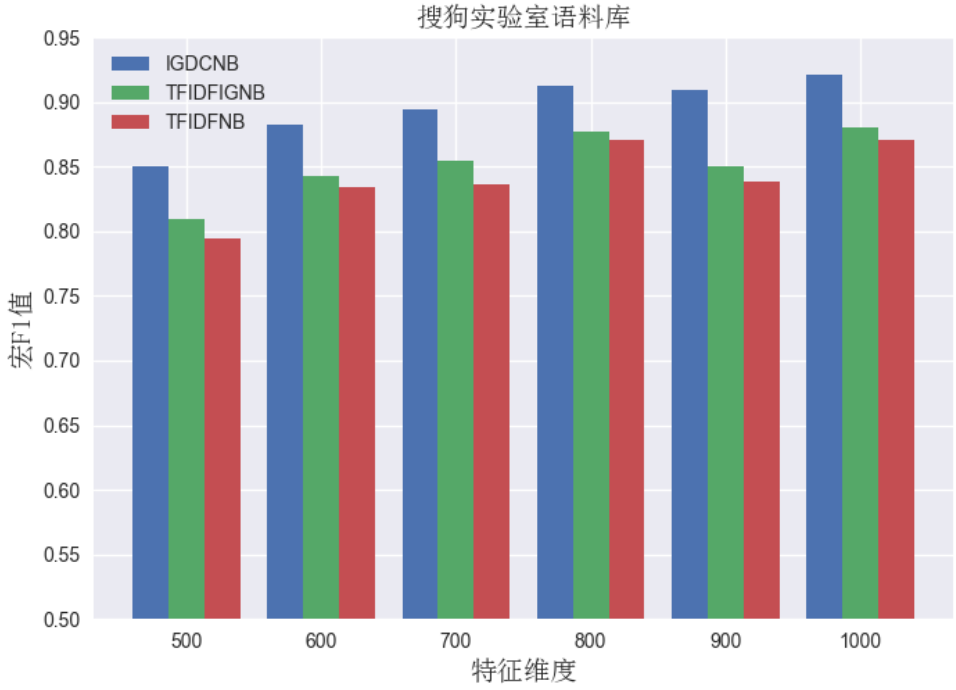

图 2.3 搜狗实验室语料库各算法宏 F1 值对比

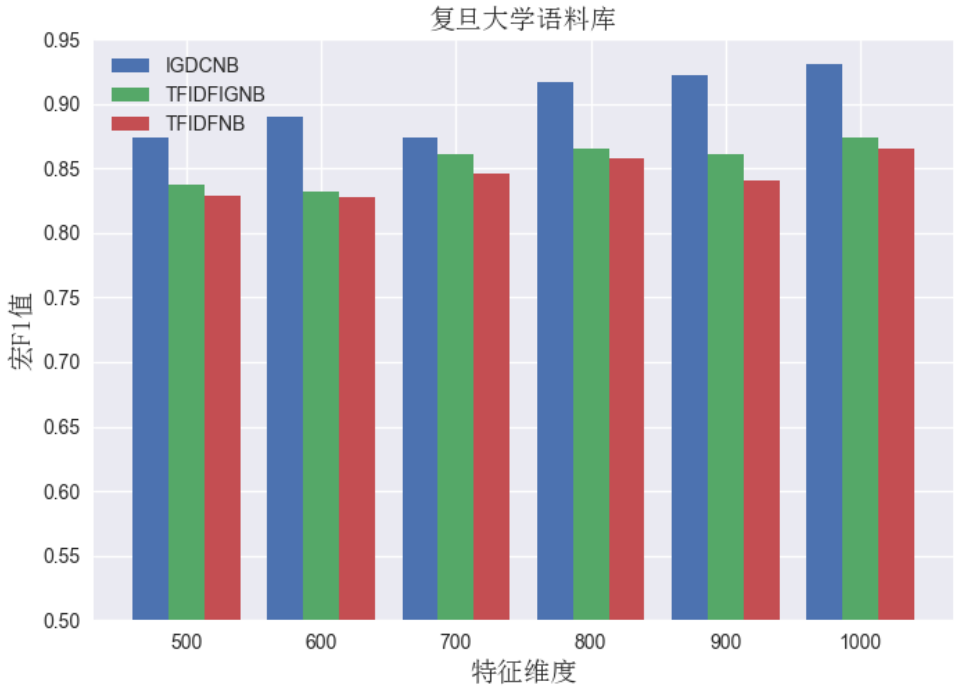

图 2.4 复旦大学语料库各算法宏 F1 值比较

从图 2.1 和 2.2 可以看出，在英文数据集上，特征维度从 500 增加到 1000 的过程中，IGDC 加权的朴素贝叶斯算法的对于文本分类的宏 F1 值都要大于 TFIDFIG 加权和 TFIDF 加权朴素贝叶斯算法，比 TFIDFIG 加权朴素贝叶斯算法高出 1%-3%，比 TFIDF 加权朴素贝叶斯算法高出 2%-4%，说明了本文算法在英文数据集上的有效性。

从图 2.3 和 2.4 可以看出，在中文数据集上，随着特征维度的增加，三个算法的宏 F1 值呈现出增长的趋势，当特征维度达到 800 时，对于两个中文数据集本文提出的 IGDC 加权朴素贝叶斯算法的宏 F1 值基本稳定在 91%左右和 92%左右，TFIDFIG 加权朴素贝叶斯算法的宏 F1 值稳定在 86%和 86%左右，TFIDF 加权朴素贝叶斯算法的宏 F1 值稳定在 85%和 85%左右；尤其当特征维度达到 800 之后，本章的 IGDC 加权朴素贝叶斯算法相比其他两个算法有了大幅的提升。综上，在中文和英文数据集上，本章的算法相较传统的算法都有了显著的提升，说明了本文算法的有效性。

2.6 本章小结

虽然朴素贝叶斯算法在文本分类中得到广泛应用，但由于其特征独立性假设的存在，导致其分类性能往往不能太好。本章中提出了一种新的加权方式：IGDC 加权，在四个国际通用的文本数据集上进行验证，结果显示本章提出的基于特征二维信息增益加权(IGDC)的朴素贝叶斯文本分类算法要比传统的 TFIDF 和 TFIDF*IG 算法都具有较高的准确率，对于各个类别的分类都更加优秀，尤其在中文数据集上，随着特征维度的增加，本章提出的算法具有更加优良的性能，因此本章提出的基于特征二维信息增益 IGDC 加权的朴素贝叶斯文本分类算法是一种有效的算法。

第三章　基于 IGDC 深度加权的朴素贝叶斯文本分类算法

上一章详细阐述了朴素贝叶斯文本分类算法的概念及模型，以及加权朴素贝叶斯的模型、常用的加权算法，并结合信息增益的优点提出了基于特征二维信息增益(IGDC)加权的朴素贝叶斯模型，实验结果表明在特征维度相同时取得了最佳的效果，而且具有更好的鲁棒性。虽然利用特征加权的方式，朴素贝叶斯文本分类算法已经得到了比较明显的优化，但是大部分学者的研究都会把加权方式放在最终的分类决策函数上，而没有考虑朴素贝叶斯算法本身的条件概率计算方式的加权，只有极少数的学者把加权方式应用于朴素贝叶斯算法中的条件概率计算方式。文献[45]将信息增益率和决策树特征加权分别作为权重用于改进朴素贝叶斯的条件概率计算公式进而构成深度加权；文献[46]将 TFIDF 作为权重，并重新改进了新的深度加权方式(DFWNB)也取得了不错的效果。

从文献[45,46]中可以看出，使用深度加权方式可以进一步提升算法的性能，本章首先详细介绍了深度加权算法的思想，并借鉴深度加权的思想提出了基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)。该算法对以往的深度加权方式进行了改进，给出了深度加权的公式，实验结果显示与上一章的 IGDCNB 算法相比，IGDC-DWNB 算法又进一步提升了朴素贝叶斯算法的性能，同时也要高于 DFWNB 算法。

3.1 特征深度加权方式

3.1.1 基于信息增益率的深度加权方法

文献[45]将信息增益率作为权重用于深度加权，若给定一个训练文档 D，特征t_k相对于训练文档的信息增益为 Gain(D,t_k)，计算公式如 3.1 所示。

$$Gain(D,t_k)=H(D)-\sum_{v\in\{0,\bar{0}\}}\frac{|D_v|}{|D|}H(D_v) \quad (3.1)$$

其中$|D_v|$表示单词w_i的取值为 v 的文档数目，$H(D)$表示训练集文档 D 的熵。

训练文档 D 关于特征词t_k的分裂信息被定义式 3.2。

$$SplitInfo(D,t_k)=-\sum_{v\in\{0,\bar{0}\}}\frac{|D_v|}{|D|}\log_2\frac{|D_v|}{|D|} \quad (3.2)$$

从而含有特征词t_k的文档 D 的信息增益率为式 3.3 所示。

$$GainRatio(D,t_k)=\frac{Gain(D,t_k)}{SplitInfo(D,t_k)} \quad (3.3)$$

进而权重的计算如式 3.4。

$$W_k=\frac{GainRatio(\mathrm{D},\mathrm{t_k})*\mathrm{m}}{\sum_{k=1}^{m}GainRatio(\mathrm{D},\mathrm{t_k})} \quad (3.4)$$

原始条件概率计算公式如式 3.5。

$$P(t_k\mid C)=\frac{\sum_{j=1}^{n}f_{jk}\delta(C_j,C)+1}{\sum_{k=1}^{m}\sum_{j=1}^{n}f_{jk}\delta(C_j,C)+m} \quad (3.5)$$

其中f_{jk}表示第 j 篇文档中的特征词t_k的频率，n 表示文档总数，m 为特征维度大小。

深度加权条件概率计算如式 3.6。

$$P(t_k\mid C)=\frac{\sum_{j=1}^{n}W_k f_{jk}\delta(C_j,C)+1}{\sum_{k=1}^{m}\sum_{j=1}^{n}W_k f_{jk}\delta(C_j,C)+m} \quad (3.6)$$

其中W_k即为上文求得的权重，$\delta(\bullet)$表示二值函数，基于信息增益率的深度加权方式如 3.7 所示。

$$c(d)=\arg\max_{c\in C}[\log_2 P(c)+\sum_{i=1}^{m}w_k f_k\log_2 p(w_i\mid c)] \quad (3.7)$$

3.1.2 基于 TFIDF 深度加权方法

文献[46]将 TFIDF 权重用于深度加权，其深度加权方式同样是改进了原始朴素贝叶斯的条件概率计算方式。TFIDF 权重的计算已经在 2.3.3 中给出，其深度加权方式定义为式 3.8。

$$c(x)=\arg\max P(c_k)\prod_{i=1}^{m}P(x_i\mid c_k,W_i)^{W_i} \quad (3.8)$$

其中W_i为第 i 个特征的权重，$P(x_i\mid c_k,W_i)^{W_i}$为深度加权条件概率，其计算方法如式 3.9。

$$P(x_i\mid c_k,W_i)^{W_i}=\frac{\sum_{i=1}^{N}W_i\delta(x_i,c_k)+1}{\sum_{i=1}^{N}W_i\delta(y_i=c_k)+n_i} \quad (3.9)$$

其中W_i即为第 i 个特征对应的 TFIDF 权重，n_i为第 i 个特征出现频数，$\delta(\bullet)$表示二值函数。

3.2 基于 IGDC 深度加权的朴素贝叶斯文本分类算法

虽然以上两种加权方式能够改善朴素贝叶斯的性能，但是由于 TFIDF 算法的缺陷不能准确的反应特征的权重，所以本章选择用 2.4 中介绍的特征二维信息增益 IGDC 作为深度加权的权重，同时朴素贝叶斯的条件概率的计算公式应与特征总数相联系，所以改进了深度加权公式。

这里中定义深度加权模型如 3.10 所示。

$$C_{map} = \max_{Cj \in C} P(C_j \mid D_i) = \max_{Cj \in C}[lnP(C_j) + \sum_{k=1}^{m} lnP(t_k \mid C_j, W_k) \times W_k \times TF_{t_k}] \qquad (3.10)$$

式中 W_k 表示特征 t_k 的权重，TF_{tk} 表示特征 t_k 在 C_j 中出现的频数，$P(t_k|C_j,W_k)$表示深度加权的特征条件概率，与[45,46]中的深度加权方式不同，本文以全新的深度加权公式 3.11 进行改进。

$$P(t_k \mid C_j, W_k) = \frac{\sum_{i=1}^{n} W_k TF_{it_k} \delta(C_i, C_j) + 1}{\sum_{k=1}^{m}\sum_{i=1}^{n} TF_{it_k}(W_k + 1)\delta(C_i, C_j) + m} \qquad (3.11)$$

式中 m 表示特征词数量，因为朴素贝叶斯的条件概率表示每个类别中出现特征的频数与该类别特征总数的比值，所以式中采用权重加一的方式平衡。式中的权重 W_k 采用上文 2.4 章节介绍二维信息增益（IGDC）加权公式，可以看出深度加权是把权重关联到朴素贝叶斯的条件概率公式中，相当于对朴素贝叶斯模型进行了改进。

定义深度加权朴素贝叶斯模型为 DWNB，将二维信息增益（IGDC）加权方式应用到深度加权朴素贝叶斯模型中得到本文模型 IGDC-DWNB，分类过程的伪代码如下表 3.1 所示。

表 3.1 算法：IGDC-DWNB

输入：训练集 D={D_1，D_2…D_i}，文档 D_i={t_1,t_2…t_k}
输出：文档标签
（1）For 文档 D_i 中的每个特征词 t_k：
　　计算每个特征词 t_k 的 *IGDC*;
（2）For 文档 D_i 的每个特征词 t_k ：
　　计算特征词 t_k 的权重 W_k;
（3）For 训练集中的文档 D_i：
　　计算先验概率 $P(C_j)$;
　　计算深度加权条件概率 $P(t_k|C_j, W_k)$;
　　计算后验概率 $P(C_j|D_i)$;
（4）选出最大的后验概率对应的类别标签，即为文档 D_i 的类别

3.3 仿真实验与分析

为了验证本章中提出的算法的分类性能，本节实现了 3.2 中提出的基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)和基于 TFIDF 深度加权的朴素贝叶斯文本分类算法(DFWNB)[46]，还有 2.4 中介绍的特征二维信息增益加权的朴素贝叶斯算法(IGDCNB)以及 TFIDF 普通加权的朴素贝叶斯文本分类算法(OFWNB)[46]。然后设计了四组实验分别在四个国际通用的文本数据集（20newsgroup，Ruster21578，搜狗实验室语料库，复旦大学语料库）上进行实验，从每个数据集中选取 60%作为训练集，40%作为测试集如表 3.2 所示。

表 3.2 数据集分布

数据集	文档数	训练集	测试集	类别数
20_newsgroup	1200	720	480	6
Ruster21578	900	540	360	6
搜狗实验室语料库	1200	720	480	6
复旦大学语料库	1200	720	480	6

在本章的实验中，分别设计四组实验，第一组实验在 20newsgroup 数据集上进行比较了 IGDC-DFWNB 与其他三个算法 DFWNB,IGDCNB,OFWNB 的性能；第二组实验在 Rusters21578 数据集上进行比较 IGDC-DWNB 与 DFWNB, IGDCNB,OFWNB 的性能；第三组实验在搜狗实验室语料库上进行验证在中文数据集上 IGDC-DWNB 与 DFWNB，IGDCNB,OFWNB 的差别；第四组实验在复旦大学语料库上验证 IGDC-DWNB 与 DFWNB, IGDCNB ,OFWNB 的性能。实验在 Python3.6，,处理器为 i5-4210，频率为 2.60GHz，内存为 8G，操作系统为 win10 的笔记本上进行。以文档频率作为选取特征词的标准每次实验进行 10 次交叉验证取平均值，在四个数据集上的查准率 P，召回率 R 以及 F1 值如下表 3.3-3.6 所示。

表 3.3 20newsgroup 上的结果，特征维度为 1000

	IGDC-DWNB			IGDCNB			DFWNB			OFWNB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
alt.atheism	0.99	0.97	**0.98**	0.96	0.97	0.97	0.86	0.97	0.92	0.98	0.93	0.96
comp.graphics	0.93	0.84	**0.89**	0.91	0.87	0.89	0.87	0.79	0.83	0.89	0.81	0.85
misc.forsale	0.77	0.96	0.85	0.82	0.93	**0.87**	0.87	0.87	0.87	0.70	0.92	0.80
rec.autos	0.94	0.89	**0.91**	0.96	0.86	0.91	0.94	0.73	0.82	0.86	0.83	0.84
sci.crypt	0.99	0.94	**0.96**	0.98	0.94	0.96	0.90	0.93	0.91	0.97	0.88	0.92
talk.politics	0.96	0.97	**0.97**	0.92	0.99	0.95	0.82	0.97	0.89	0.94	0.95	0.94
平均值	**0.93**	**0.93**	**0.93**	0.92	0.92	0.92	0.87	0.87	0.87	0.89	0.88	0.88

表 3.4 Ruster21578 上的结果，特征维度为 1000

	IGDC-DWNB			IGDCNB			DFWNB			OFWNB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Crude	0.97	0.92	**0.94**	0.90	0.96	0.93	0.95	0.80	0.87	0.98	0.82	0.89
Acq	0.92	0.97	**0.94**	0.90	0.98	0.94	0.72	0.95	0.82	0.72	0.95	0.82
Interest	0.64	0.87	**0.74**	0.64	0.75	0.69	0.67	0.58	0.62	0.61	0.79	0.69
Wheat	0.97	1.00	**0.99**	0.96	1.00	0.98	0.98	0.94	0.96	0.95	0.98	0.96
Money	0.81	0.49	0.61	0.72	0.56	0.63	0.61	0.73	**0.66**	0.73	0.49	0.59
Earn	0.92	0.95	**0.93**	0.99	0.89	0.94	0.97	0.81	0.88	0.95	0.84	0.89
平均值	**0.87**	**0.87**	**0.87**	0.85	0.85	0.85	0.81	0.80	0.80	0.82	0.81	0.80

表 3.5 搜狗实验室语料库上的结果，特征维度为 1000

	IGDC-DWNB			IGDCNB			DFWNB			OFWNB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
健康	0.90	0.82	**0.86**	0.93	0.73	0.82	0.95	0.66	0.78	0.89	0.72	0.79
教育	0.82	0.87	**0.85**	0.78	0.93	0.85	0.82	0.84	0.83	0.79	0.79	0.79
军事	0.96	0.99	**0.97**	0.93	0.99	0.95	0.89	0.93	0.91	0.85	0.94	0.89
旅游	0.89	0.94	**0.91**	0.89	0.94	0.91	0.91	0.87	0.89	0.88	0.89	0.88
体育	0.91	0.95	**0.93**	0.94	0.93	0.93	0.98	0.93	0.95	0.83	0.87	0.85
文化	0.88	0.83	**0.86**	0.85	0.84	0.85	0.66	0.89	0.76	0.75	0.80	0.77
平均值	**0.89**	**0.90**	**0.90**	0.88	0.89	0.88	0.86	0.85	0.85	0.83	0.83	0.83

表 3.6 复旦大学语料库上的结果，特征维度为 1000

	IGDC-DWNB			IGDCNB			DFWNB			OFWNB		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
环境	0.96	0.89	**0.92**	0.95	0.86	0.90	0.87	0.87	0.87	0.94	0.84	0.89
交通	0.95	0.95	**0.95**	0.93	0.93	0.93	0.95	0.80	0.86	0.86	0.84	0.85
教育	0.97	0.96	**0.97**	0.93	0.97	0.95	0.95	0.96	0.95	0.96	0.93	0.95
军事	0.94	0.87	0.90	0.93	0.93	**0.93**	0.91	0.82	0.86	0.94	0.76	0.84
经济	0.94	0.99	**0.96**	0.93	0.94	0.94	0.85	0.90	0.87	0.80	0.85	0.83
体育	0.88	0.98	0.93	0.95	0.97	**0.96**	0.83	0.98	0.90	0.75	0.98	0.85
平均值	**0.94**	**0.94**	**0.94**	0.93	0.93	0.93	0.89	0.89	0.89	0.87	0.86	0.86

从表 3.3 和表 3.4 中可以看出在英文数据集上本文提出的深度加权方式要比 2.4 节中提出的加权方式更加有效，这是因为深度加权是把权重关联到了朴素贝叶斯分类器的条件概率计算公式上，实际上是对朴素贝叶斯模型进行了改进，不仅让每个特征都具有不同的权重，而且改进了分类器本身。一方面本文提出的改进的深度加权方式比 2.4 节中的单纯加权方式总体提升了 1%到 2%，另一方面可以看出论文[46]提出的深度加权方式在英文数据集上对于每个类别的提升都不是太大，总体上反而要比一般的加权方式低一些。本文的深度加权方式要

比 DFWNB 在每个类别上都要平均高出 6%左右，比 OFWNB 算法高出 5%左右，说明了本文的改进深度加权朴素贝叶斯文本分类算法在英文数据集上的有效性。

从 3.5 和表 3.6 中可以看出在中文数据集上，总体上对于所有类别的 F1 平均值本文的 IGDC-DWNB 要大于 IGDCNB 算法，IGDCNB 算法要大于 DFWNB 算法，DFWNB 算法又要大于 OFWNB 算法。一方面显示了本文的改进深度加权算法在原始加权算法上的有效性，另一方面虽然论文[46]提出的深度加权算法在英文数据集上的并没有取得较好的效果，但是在中文数据集上确实要优于原始的 TFIDF 普通加权算法。进一步观察可以发现在搜狗实验室数据集上本文的 IGDC-DWNB 算法在每个类别上都取得了最好的成绩，比 IGDC 加权朴素贝叶斯算法平均高出 2%，比 TFIDF 深度加权朴素贝叶斯算法平均高出 5%，比 TFIDF 普通加权朴素贝叶斯算法平均高出 7%左右。而在复旦大学语料库上，本文的 IGDC-DWNB 算法虽然在经济和体育两个类别上没有超过 IGDCNB 算法，但整体上要比 IGDCNB 算法高出 1%，比 DFWNB 高出 5%，比 OFWNB 高出 7%左右，也充分显示了文本的改进深度加权的朴素贝叶斯文本分类算法在中文数据集上的有效性。

为了比较算法在对于整个语料库的分类性能，我们计算了所有类别对应的宏 F1 值，并在六个不同特征维度下进行了仿真实验，在四个数据集上六个不同维度下的类别宏 F1 值对比如下图 3.1-3.4 所示。

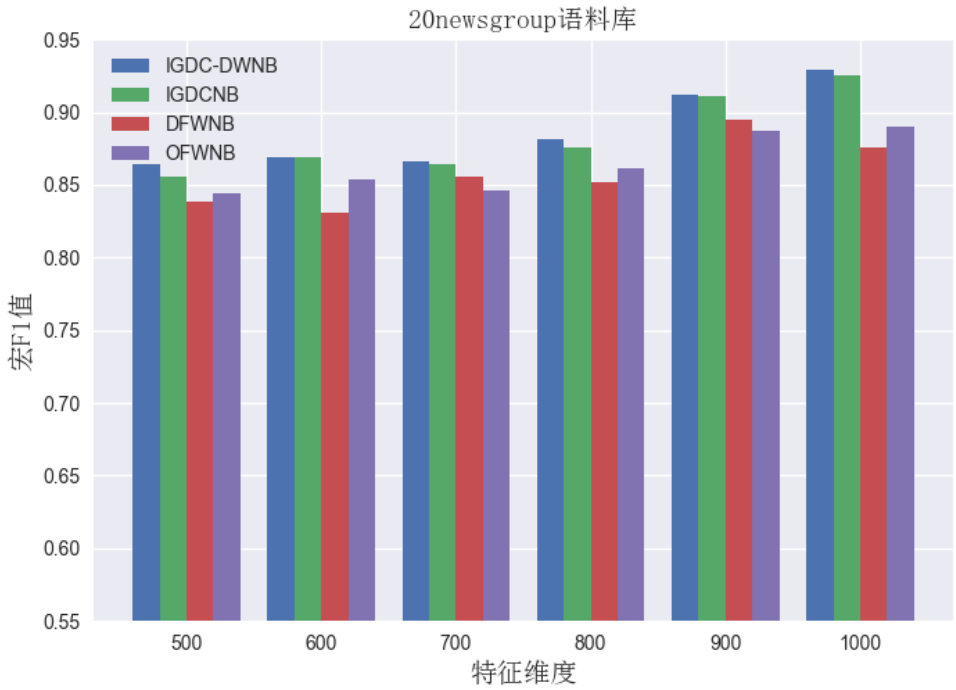

图 3.1　20Newsgroup 语料库各算法宏 F1 值对比

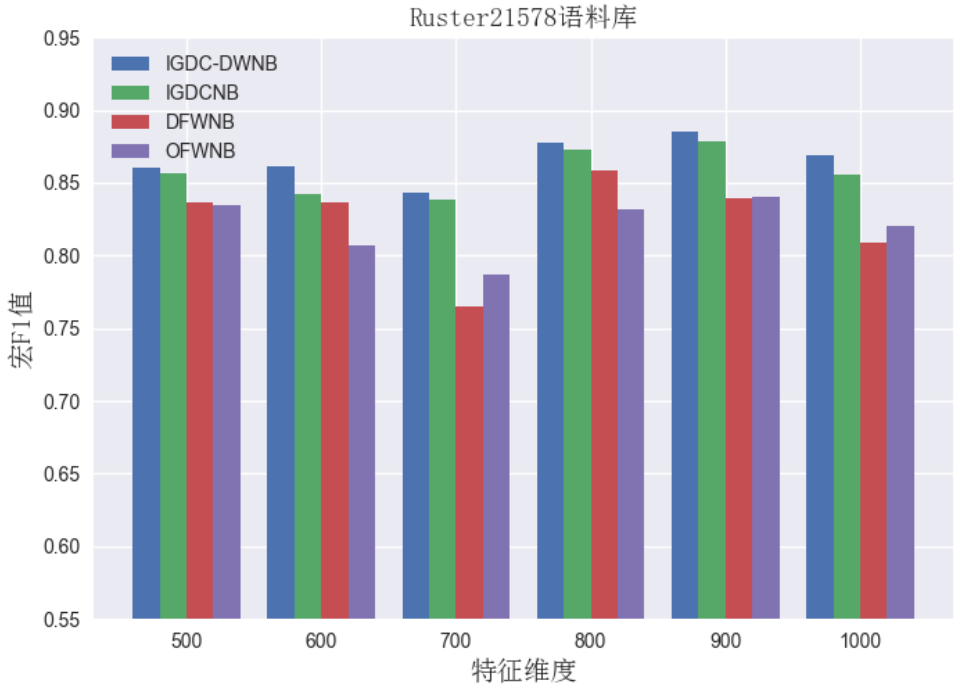

图 3.2　Ruster21578 语料库各算法宏 F1 值对比

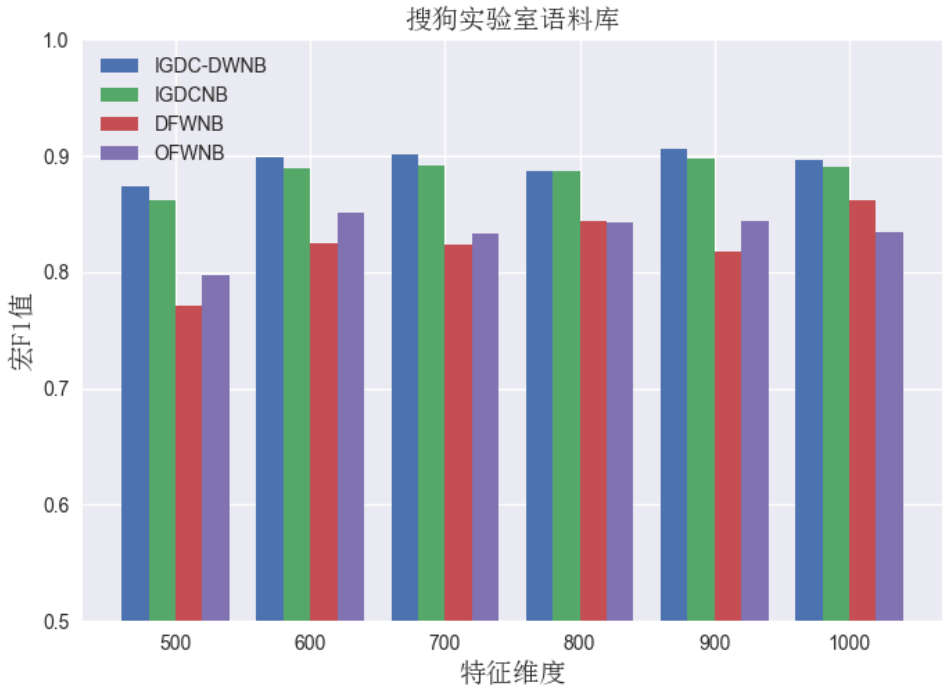

图 3.3 搜狗实验室语料库各算法宏 F1 值对比

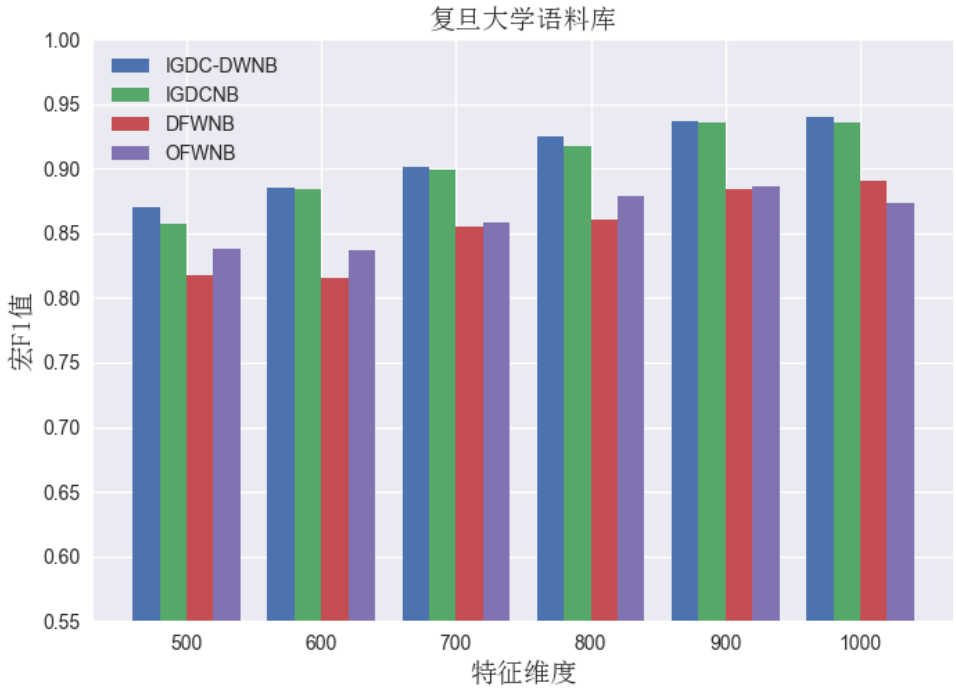

图 3.4　复旦大学语料库各算法宏 F1 值对比

从图 3.1 和 3.2 可以看出，在 20newsgroup 数据集上随着特征维度的增加，宏 F1 值总体呈现上升的趋势，在六个维度下本文的 IGDC-DWNB 算法的宏 F1 值都要高于其他算法，其中在特征维度为 500,800,1000 时提升幅度比较大，TFIDF 深度加权算法 DFWNB 只在两个维度下优于 OFWNB，算法的鲁棒性要弱于 IGDC-DWNB 算法。在 Rusters21578 数据集上，随着特征维度的增加，IGDC-DWNB 算法的宏 F1 值表现的比较平稳，DFWNB 算法的波动比较大，说明了其算法鲁棒性不好。在六个特征维度下 IGDCNB 算法也要弱于 IGDC-DWNB，IGDC-DWNB 算法比普通加权平均提升了 1%到 2%，说明文本提出的改进深度加权的在英文数据集上的有效性。

从图 3.3 和 3.4 中可以看出，在搜狗实验室语料库中改进深度加权的朴素贝叶斯文本分类算法 IGDC-DWNB 在不同维度下的宏 F1 值也是最高的，比 IGDC 普通加权算法高出 2%左右。同时 TFIDF 深度加权朴素贝叶斯算法(DFWNB)在维度较低时并没有普通加权算法好，在维度达到 1000 时要优于普通加权。在复旦大学语料库中，IGDC-DWNB 算法的宏 F1 值也要高于其他三个算法，在两个中文数据集下，DFWNB 算法的相同点是低纬度下该算法性能要弱于普通加权，高纬度下优于普通加权。以上在英文和中文数据集下的对比可知，本章的改进深度加权的朴素贝叶斯文本分类算法具有较强的鲁棒性和有效性，深度加权方法要优于普通加权方法。

3.4 本章小结

尽管普通加权朴素贝叶斯算法在文本分类上能够取得不错的效果，但由于其只把权重单纯的与贝叶斯决策函数关联并没有把提升最大化，所以文献[46]把 TFIDF 权重加入到贝叶斯条件概率计算公式中构成深度加权，但其加权方式忽略了特征数量与总特征数量之间的关系，导致算法的鲁棒性比较差，不能在大多数数据集上有提升效果。同时，由于朴素贝叶斯的条件概率表示每个类别中出现特征的频数与该类别特征总数的比值，而文献[45]中的深度加权方式忽略了这一点，所以本章中采用权重加一的方式平衡，提出了改进深度加权的朴素贝叶斯文本分类算法（IGDC-DWNB），并在四个国际通用的中英文本数据集上进行实验研究，结果显示本文提出的基于 IGDC 深度加权方式是对朴素贝叶斯模型进行了改进，相比于其他三种算法具有更佳的分类性能，分类效果的提升更大，与上一章提出的 IGDCNB 相比也有很大的提升，因此，本章提出的基于 IGDC 深度加权的朴素贝叶斯文本分类算法是有效的算法。

第四章　改进的自定义特征维度的快速相关性过滤算法

第三章在加权的基础上对朴素贝叶斯的模型进行了改进，提出了基于 IGDC 深度加权的朴素贝叶斯算法(IGDC-DWNB)，实验结果表明新提出的 IGDC-DWNB 算法与文献[46]的 DFWNB 相比在特征维度相同时具有更高的精确度，提升了算法的分类性能。另一方面，特征选择也是影响文本分类效果的重要因素之一，为了进一步提高算法的性能，本章从特征选择入手对朴素贝叶斯算法进行改进。

特征选择在文本数据预处理中是一项非常重要的工作，在文本分类任务中现有的特征选择算法大部分都没有考虑特征间的冗余性。在特征选择时能够去除冗余特征时非常重要的，不仅能够降低模型的时间复杂度，而且在一定程度上能够提升分类器的性能。机器学习中主要有两种特征选择方法：过滤器和包装器。过滤方法选择一个特征子集作为预处理步骤，它独立于分类算法工作。包装法具有更高的复杂度在选择特征时也需要更多的时间，对于文本分类任务显然是不可取的[49]。因此我们把关注点聚焦在过滤方法上。本章中，为了去除特征冗余性，本章借鉴了文献[56]中的快速相关性过滤算法 FCBF 的优点，后续的实验发现原始的 FCBF 算法并不适用于文本分类任务，所以本章改进了原始的快速过滤性算法(FCBF)特征相关性的计算公式并完善了算法流程，提出了改进的自定义特征维数快速相关性过滤特征选择算法，我们简称为（IFSC-FCBF）。在仿真实验中，我们将 IFSC-FCBF 与朴素贝叶斯分类算法相结合，并在四个通用语料数据库中进行验证，与其他特征选择算法相比结果显示我们的算法在相同的特征维数下，能具有更高的准确率，同时具有更低的时间复杂度，能够更加有效的去除冗余特征。

4.1 过滤式文本特征选择方法

4.1.1 信息增益特征选择算法(IG)

要信息增益显示了一个特征对于文档分类正确与否的贡献率[13]，即如果一个特征词拥有比较大的信息增益，那么它对文档的分类作用就越大，特征词 t 所具有的信息增益可由公式 4.1 计算得到。

$$IG(t) = -\sum_{j=1}^{M} P(C_j)\log P(C_j) + \sum_{t}\left|\frac{Dt_k}{D}\right|\sum_{j=1}^{M} P(C_j \mid t_k)\log P(C_j \mid t_k) \qquad (4.1)$$

其中 M 表示类别数， $P(C_j)$ 表示类别 C_j 出现的概率， Dt_k 表示含有特征词 t_k 的文档数， D 表示总的文档数， $P(C_j|t_k)$ 表示给定特征词 t_k 后类别 C_j 出现条件概率

4.1.2 判别式特征选择算法(DFS)

DFS 是最近比较成功的特征选择算法[16]，算法思想是在给定一个类别时选择一些有区别性的特征并消除对分类没有作用的特征。理想的特征选择方法应该是为区别性的特征分配较高的分数，为冗余的特征分配较低的分数。对于文本分类，每个区别性的词对应一个特征，而每个特征应该具有如下四个要求：

1.一个特征，经常出现在一个单独的类中，而不会出现在另一个类中，那么这个特征具有较强的区别性; 因此，必须给它一个高分。

2.一个特征如果很少出现在单个类别中而且不在其他类别中出现是无关紧要的; 因此，需要分配低分。

3.一个特征如果经常出现在所有类别中，那么也是无关紧要的; 因此，它一定是分配低分。

4.在一些特定类别中经常出现的特征具有较强的区别性，因此，它必须分配相对较高的分数。

根据前两条的要求，特征得分的计算可由公式 4.2 计算得到。

$$\sum_{i=1}^{M}\frac{P(C_i|t_k)}{P(\overline{t_k}|C_i)+1} \qquad (4.2)$$

其中 M 是所有的类别数， $P(C_i|t_k)$ 表示具有特征 t_k 的类别 C_i 出现的概率， $P(\overline{t_k}|C_i)$ 表示不含有 t_k 的类别 C_i 出现的概率。从上式中可以明显看出如果一个特征词在一个类的所有文档中都出现过而在其他类中的文档中没有出现过将被指定为最高分 1.0。如果一个特征词在一个类中少出现，尽管没有在其他类别中出现也将获得较低的分数。此时这个公式还不满足第三种要求，因为所有类的每个文档中都出现的特征也会获得最高分数 1.0。 为了解决这个问题，该公式扩展为 4.3 所示。

$$DFS(t_k)=\sum_{j=1}^{M}\frac{P(C_j|t_k)}{P(\overline{t_k}|C_j)+P(t_k|\overline{C_j})+1} \qquad (4.3)$$

其中 M 表示类别数， $P(C_j|t_k)$ 表示给定特征词 t_k 后类别 C_j 出现条件概率， $P(\overline{t_k}|C_j)$ 表示类别 C_j 中除特征词 t_k 外其他特征词出现的条件概率， $P(t_k|\overline{C_j})$ 表示除类别 C_j 外其他类别中特征词 t_k 出现的条件概率。

4.1.3 基于相关性的快速过滤(FCBF)算法

基于相关性的快速过滤算法（FCBF）是一种快速过滤的特征选择算法，使用对称的不确定性来度量两个特征的相关性，是一种典型的启发式序列后向消除方法[56]。基于相关性的快速过滤算法的思想是分别计算每个特征与类别之间的相关性，特征之间的互相关性，如果特征与类别之间的相关性大于特征与特征之间的相关性则保留，否则剔除，其选择过程如下：首先计算变量 X 的信息熵如 4.4 所示。

$$H(X)=-\sum_i P(x_i)\log P(x_i) \qquad (4.4)$$

其中 $P(x_i)$ 表示变量 X 取值为 x_i 时的概率，在变量 Y 出现时变量 X 的条件信息熵定义为 4.5。

$$H(X\mid Y)=-\sum_j P(y_j)\sum_i P(x_i\mid y_i)\log P(x_i\mid y_i) \qquad (4.5)$$

其中 $P(y_j)$ 表示变量 Y 取值为 y_j 时的概率，$P(x_i\mid y_i)$ 表示给定变量 Y 时变量 X 的条件概率，当变量 Y 出现时变量 X 信息熵的变化量即为信息增益，如式 4.6 所示。

$$IG(X\mid Y)=H(X)-H(X\mid Y) \qquad (4.6)$$

于是对称不确定性可由式 4.7 计算得到。

$$SU(X,Y)=2\left(\frac{IG(X\mid Y)}{H(X)+H(Y)}\right) \qquad (4.7)$$

FCBF 的思想在于，如果一个特征与类别之间的对称不确定性比较高，而与其他特征之间的对称不确定性比较低时，认为该特征即为被选中的显著特征。首先计算每个特征与类别间的对称不确定性 $SU_{f_i,c}$，并以 SU 最大值对应的特征作为显著特征 f_q，分别计算其他特征 f_p 与显著特征的对称不确定性 SU_{f_q,f_p}，如果 $SU_{f_q,f_p}\geq SU_{f_p,c}$ 则移除特征 f_p，并在剩余的特征中选出显著特征重复上述步骤，最后得到的特征子集都是有显著特征组成。

4.2 改进的自定义维度的快速相关性过滤算法

尽管 FCBF 在众多领域取得好成绩，但是下一节的实验显示在文本分类任务中它的表现并不是很理想，究其原因我们发现是由于原始 FCBF 算法在计算信息增益时没有考虑特征在文本中的分布情况，于是我们采用特征的分布信息计算信息增益，进而计算相关性.另一方面 FCBF 算法会出现特征与特征之间的相关性会比较高，而特征与类别之间的相关性比较低的情况，这导致能够保留下来的特征很少，影响分类精度[72]，因此我们加入了特征维度参数，

能够自定义特征维度，使特征选择更加平衡，同时改进了原始 FCBF 的相关性的计算方法。

特征与类别的相关性计算：

设文档类别为 $C=\{C_1,C_2...C_j\}$，$j=1,2,3...V$，设 D_i 为任意一篇训练文档，$D_i=\{tf(t_1),tf(t_2)....tf(t_m)\}$。信息熵是反应一个变量不确定性程度的物理量，在文本分类中信息熵反应的就是变量在语料库中的分布均匀的程度，对于类别和特征来说，其信息熵可定义分别为式 4. 8, 4. 10。

$$H(C)=-\sum_{j=1}^{V}P(C_j)\times\log_2 P(C_j) \quad (4.8)$$

$$P(C_j)=\frac{\sum_{i=1}^{n}\delta(C_i,C_j)+L}{n+V\times L} \quad (4.9)$$

$$H(t_k)=-\sum_{j=1}^{V}P(t_k)\times\log_2 P(t_k) \quad (4.10)$$

$$P(t_k)=\frac{tf(t_{k,i})+L}{\sum_{i=1}^{n}tf(t_{k,i})+V\times L} \quad (4.11)$$

式中 n 表示训练文档总数，V 表示类别个数，C_i 表示训练文档 D_i 的类别，$P(t_k)$ 表示特征词 t_k 出现的概率，$tf(t_{k,i})$ 表示特征词 t_k 在训练文档 D_i 出现的频数，L 是为了抑制概率为 0 的情况所加入的平滑因子，本文中取 L =0. 001，$\delta(\bullet)$ 表示二值函数其公式如式 4.12。

$$\delta(x,y)=\begin{cases}1, x=y\\0, x\neq y\end{cases} \quad (4.12)$$

对于单个特征来说，在已经类别 C_j 分布的情况下的条件信息熵为 4.13 所示。

$$H(t_k\mid C)=-\sum_{j=1}^{V}P(t_k\mid C_j)\times\log_2 P(t_k\mid C_j) \quad (4.13)$$

$$P(t_k\mid C_j)=\frac{tf(t_k\mid C_j)+L}{\sum_{j=1}^{V}tf(t_k\mid C_j)+V\times L} \quad (4.14)$$

式中 $P(t_k\mid C_j)$ 表示已经类别 C_j 时，特征词 t_k 出现的概率，$tf(t_k\mid C_j)$ 表示特征词 t_k 在 C_j 类中的频数，L 是为了抑制概率为 0 的情况所加入的平滑因子，本文中取 L =0.001，V 表示类别数。

在已经类别的分布情况下特征 t_k 信息熵的变化量即为特征信息增益[73]，计算公式如 4.15 所示。

$$IG(t_k\mid C)=H(t_k)-H(t_k\mid C) \quad (4.15)$$

由此特征 t_k 与类别的相关性可由式 4.16 计算得到。

$$Corr(t_k,C)=\frac{IG(t_k\mid C)}{\sqrt{H(t_k)}\times\sqrt{H(C)}} \qquad (4.16)$$

特征与特征的相关性：

对于t_{k1}来说，在已知t_{k2}在各类别中的分布情况下的条件信息熵如式 4.17 所示。

$$H(t_{k1}\mid t_{k2})=-\sum_{j=1}^{V}P(t_{k1}\mid t_{k2})\log_2 P(t_{k1}\mid t_{k2}) \qquad (4.17)$$

$$P(t_{k1}\mid t_{k2})=\frac{df(t_{k1},t_{k2}\mid C_j)}{df(t_{k2}\mid C_j)} \qquad (4.18)$$

其中$P(t_{k1}\mid t_{k2})$表示特征t_{k2}出现时t_{k1}也出现的概率，$df(t_{k1},t_{k2}\mid C_j)$表示在类别$C_j$中$t_{k1}$和$t_{k2}$同时出现的文档数，$df(t_{k2}\mid C_j)$表示在类别$C_j$中特征词$t_{k2}$出现的文档数。

对于是否存在t_{k2}，特征t_{k1}在类别C_j中信息熵的变化量即为式 4.19,。

$$IG(t_{k1}\mid t_{k2})=H(t_{k1}\mid C)-H(t_{k1}\mid t_{k2}) \qquad (4.19)$$

t_{k1}和t_{k2}之间的相关性可由式 4.20 计算得到。

$$Corr(t_{k1},t_{k2})=\frac{IG(t_{k1}\mid t_{k2})}{\sqrt{H(t_{k1})}\times\sqrt{H(t_{k2})}} \qquad (4.20)$$

为了保证 IFSC-FCB 特征选择算法能够得到自定义维度的特征，我们对算法的流程也进行了改进，需要进行两种情况的判断：当所有特征都经过筛选后，最终输出的特征列表的维度小于我们设定的维度时，我们认为特征与类别的相关性作为显著相关性，这意味着我们更看重$Corr(t_k,C)$的值，我们选择出具有较大$Corr(t_k,C)$值的特征去补足；当最终输出的特征列表刚好大于或等于我们设定的特征维度时，可以直接从最终输出取到足够的特征。

IFSC-FCBF 算法流程如下表 4.1 所示。

表 4.1 IFSC-FCBF 算法流程

输入：向量化文本矩阵X，类别标签矩阵C
输出：特征列表S_{best}
1.初始化$T=\{t_1,t_2....t_m\}$，$S_{list}=\{\}$,size 是个整数，thresh 是个小数
2.对$t_k\in T$，计算t_k与类别的相关性$Corr(t_k,C)$，$Corr(t_k,C)\geq thresh$时添加t_k进S_{list}
3.将S_{list}按照$Corr(t_k,C)$值从大到小排序，$t_p=getFirst(S_{list})$，$S_{best}=\{t_p\}$
4. do
　　$t_q=getNext(S_{list})$，
　　if $t_q\neq NULL$
　　　　计算$Corr(t_p,t_q)$
　　　　if($Corr(t_p,t_q)\geq$)$Corr(t_q,C)$
　　　　　　remove t_q from S_{list}

else
append t_q to S_{best}
$t_q = getNext(S_{list})$
if $length(S_{best}) \geq size$
return S_{best}
else $t_p = getNext(S_{list})$
until $t_p == NULL$
if $length(S_{best}) < size$
$S_{best} = S_{best} + S_{list}[:(size - length(S_{best}))]$
return S_{best}

该算法在复杂度方面是在 FCBF 算法基础上增加了两个判断，所以复杂度与 FCBF 算法相同[56]，最大时间复杂度仍然为$O(MN\log N)$，N 是预先选出的特征数目，M 是需要计算的特征对数。

4.3 实验设计与结果分析

为了验证本文提出的改进的自定义特征维度的快速相关性过滤算法的有效性，首先实现了本章提出的改进的自定义特征维度的相关性快速过滤算法（IFSC-FCBF），信息增益(IG)过滤算法[51],区别性特征选择算法(DFS)[55]，以及基于相关性快速过滤算法（FCBF）[56],并把上述四个算法与普通朴素贝叶斯分类器相结合构成分别在四个数据集 20newsgroup, Ruster21578，搜狗实验室语料库和复旦大学中文语料库上进行四组实验。实验在 Python3.6，,Ubuntu Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz，125G 的 RAM 环境下进行，选取 60%作为训练集，40%作为测试集进行验证。实验数据分布如下表 4.2 所示。

表 4.2 数据集分布

数据集	文档数	训练集	测试集	类别数
20_newsgroup	1200	720	480	6
Ruster21578	900	540	360	6
搜狗实验室语料库	1200	720	480	6
复旦大学语料库	1200	720	480	6

首先在英文数据集上进行实验，为了观察选择的特征数对宏 F1 值的影响，我们把选择的特征数从 100 不断增加到 500，并比较了上述四个算法的性能，如图 4.1-4.2 所示。

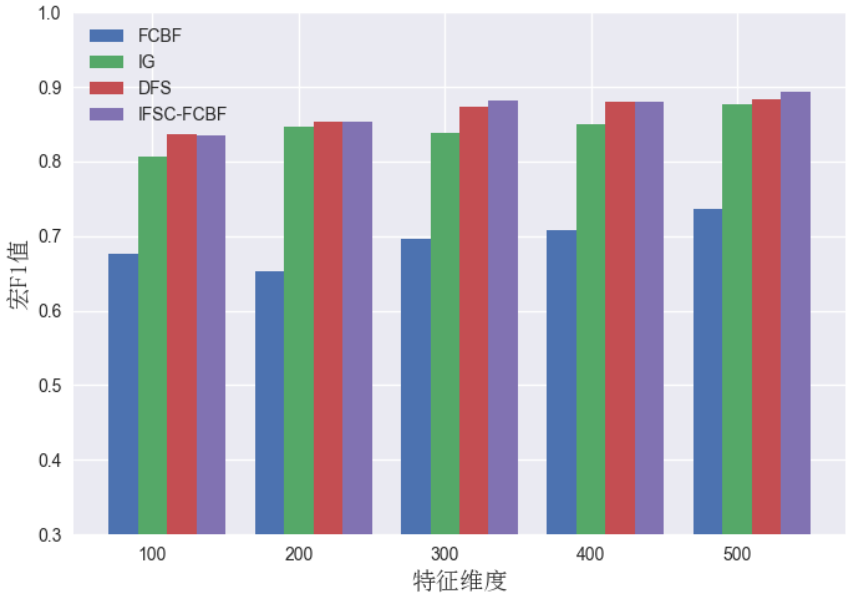

图 4.1:20newsgroup 数据集的算法性能比较

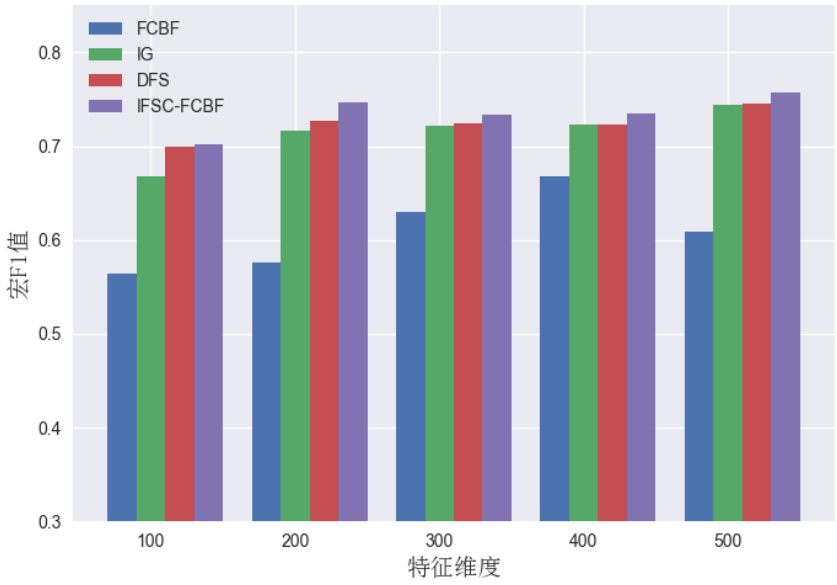

图 4.2：Ruster21678 数据集的算法性能比较

由图 4.1 和图 4.2 可以看出，在两个英文数据集上，随着特征数量的增加，宏 F1 值会有略微的增加，在特征数量到达 300 时，宏 F1 值趋于平稳。我们可以看到两个图的共同点，本章提出的 IFSC-FCBF 特征选择算法能够更加有效的选择特征，获得了最大的宏 F1 值。一方面，由前部分内容介绍可知，由于原始的 FCBF 算法会出现消除特征过剩的情况，所以在文本分类中并不能得到很好的结果，在四个算法中其性能是最差的。另一方面，DFS 特征选择算法要优于 IG 算法，特别是在 20newsgroup 数据集上。可以看出在特征数目在 300 时，DFS 的 F1 值要比 IG 的 F1 值高出 3%左右。虽然在 20newsgroup 数据集上，IFSC-FCBF 算法结果

和 DFS 很接近甚至有时候会落后于 DFS，但在 Ruster21578 上，IFSC-FCBF 的性能一直都是优于 DFS。总的来说，在英文数据集上，IFSC-FCBF 比他的对手具有更好的性能，宏 F1 值比 DFS 算法平均高出 1%，比 IG 算法平均高出 2%到 3%。

在中文数据集上，我们运用同样的方式得到的结果如图 4.3-4.4 所示。

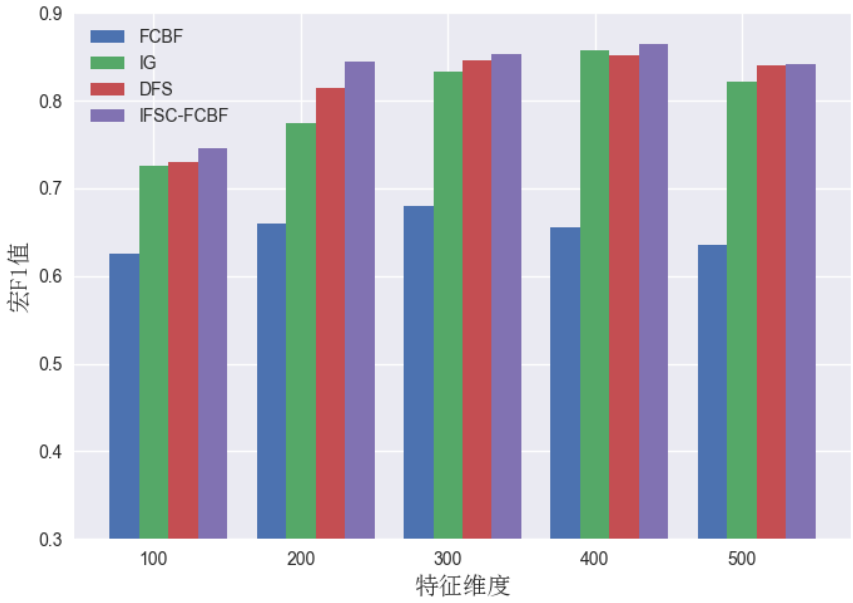

图 4.3：复旦语料库的算法性能比较

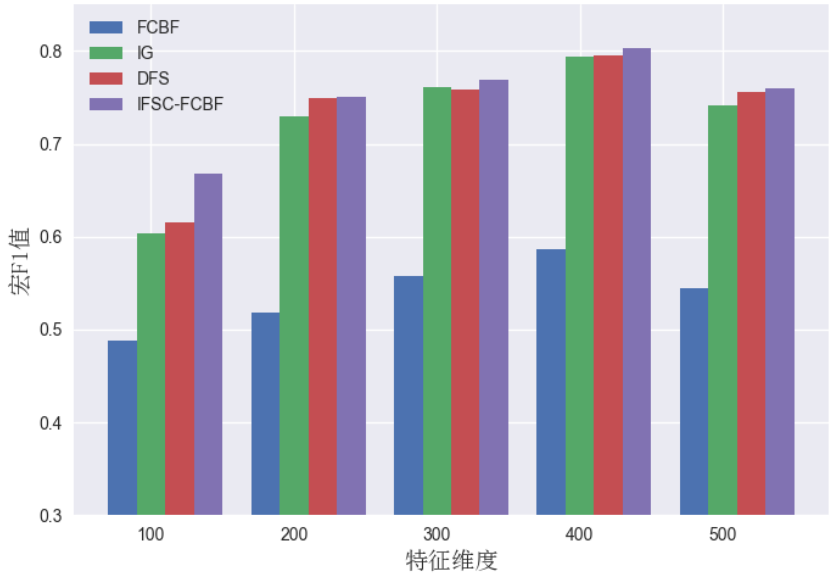

图 4.4：搜狗语料库的算法性能比较

由图 4.3 和图 4.4 可以看出，随着特征数增加到 300 时，宏 F1 值基本趋向平稳，与在英文数据集上的结果类似。这意味着当特征维数达到 300 时，特征已经不再是影响算法性能的因素。对于 FCBF 算法来说，在英文数据集和中文语料库上的表现都不是很好。DFS 算法在

在复旦语料库中宏 F1 值比 IG 算法平均高出 1.4%，在搜狗语料库中平均高出 0.8%。而 IFSC-FCBF 算法比 DFS 平均高出 1.3%和 1.5%。

为了看出各特征选择算法对每个类别的分类效果，我们在特征维数为 300 时，对每个类别的分类效果做了统计，如下表 4.3-4.6 所示。

表 4.3 20newsgroup 数据集各类别分类效果比较

	IFSC-FCBF			DFS			IG			FCBF		
类别	P	R	F1	P	R	F1	P	R	F1	P	R	F1
alt.atheism	0.95	0.86	**0.91**	0.94	0.84	0.88	0.63	0.99	0.77	0.44	0.79	0.56
comp.graphics	0.82	0.82	0.82	0.90	0.82	**0.86**	0.89	0.73	0.80	0.78	0.66	0.72
misc.forsale	0.98	0.94	**0.96**	0.93	0.94	0.94	0.96	0.97	0.96	0.66	0.84	0.74
rec.autos	0.95	0.94	0.94	0.95	0.95	**0.95**	0.94	0.96	0.95	0.97	0.78	0.86
sci.crypt	0.84	0.81	**0.83**	0.82	0.80	0.81	0.83	0.70	0.76	0.94	0.56	0.70
talk.politics.guns	0.76	0.90	**0.82**	0.73	0.87	0.80	0.90	0.71	0.79	0.71	0.51	0.60
平均值	**0.89**	**0.88**	**0.88**	0.88	0.87	0.87	0.86	0.84	0.84	0.76	0.69	0.70

表 4.4 Ruster21578 数据集各类别分类效果比较

	IFSC-FCBF			DFS			IG			FCBF		
类别	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Crude	0.61	0.62	0.62	0.58	0.57	0.57	0.60	0.55	0.58	0.75	0.66	**0.70**
Acq	0.65	0.45	0.53	0.70	0.43	0.53	0.62	0.49	**0.55**	0.25	0.26	0.26
Interest	1.00	0.96	**0.98**	1.00	0.94	0.97	1.00	0.95	0.97	0.97	0.83	0.89
Wheat	1.00	1.00	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.90	0.87
Money	0.59	0.80	0.68	0.60	0.87	**0.71**	0.59	0.75	0.66	0.48	0.67	0.56
Earn	0.60	0.58	**0.59**	0.56	0.57	0.56	0.54	0.60	0.57	0.56	0.45	0.50
平均值	**0.76**	**0.76**	**0.76**	0.76	0.75	0.75	0.75	0.74	0.74	0.67	0.65	0.66

表 4.5 复旦大学语料库各类别分类效果比较

	IFSC-FCBF			DFS			IG			FCBF		
类别	P	R	F1	P	R	F1	P	R	F1	P	R	F1
环境	0.95	0.77	0.85	0.93	0.82	0.87	0.94	0.84	**0.89**	0.68	0.18	0.29
交通	0.68	0.91	0.78	0.91	0.79	**0.85**	0.98	0.58	0.73	0.83	0.47	0.60
教育	0.95	0.85	**0.90**	0.91	0.87	0.89	0.95	0.84	0.89	0.61	0.72	0.66
军事	0.96	0.88	0.92	0.70	1.00	0.82	0.99	0.88	**0.93**	0.61	0.72	0.66
经济	0.81	0.85	0.83	0.85	0.90	**0.88**	0.88	0.88	0.88	0.96	0.59	0.73
体育	0.81	0.82	**0.81**	0.87	0.72	0.79	0.48	0.91	0.63	0.83	0.32	0.46
平均值	**0.86**	**0.85**	**0.85**	0.86	0.85	0.85	0.88	0.82	0.83	0.24	0.84	0.37

表 4.6 搜狗实验室语料库各类别分类效果比较

	IFSC-FCBF			DFS			IG			FCBF		
类别	P	R	F1	P	R	F1	P	R	F1	P	R	F1
健康	0.92	0.80	**0.86**	0.86	0.83	0.85	0.81	0.80	0.81	0.85	0.46	0.60
教育	0.92	0.83	**0.87**	0.91	0.82	0.86	0.91	0.83	0.87	0.96	0.59	0.73

军事	0.50	0.90	0.64	0.60	0.66	0.64	0.52	0.88	**0.65**	0.26	0.76	0.38
旅游	0.72	0.64	**0.68**	0.62	0.66	0.64	0.78	0.59	0.67	0.48	0.57	0.52
体育	0.90	0.83	0.86	0.89	0.85	**0.87**	0.90	0.84	0.87	0.92	0.51	0.66
文化	0.84	0.60	0.70	0.70	0.70	0.70	0.79	0.63	**0.76**	0.78	0.32	0.46
平均值	**0.81**	**0.76**	**0.77**	0.77	0.76	0.76	0.79	0.76	0.76	0.72	0.53	0.56

表 4.3 显示的是特征选择算法在 20newsgroup 上对每个类别的分类效果的比较，如表格中的粗体所示，在大部分类别中 IFSC-FCBF 都具有最高的 F1 值，其次是 DFS 算法。因为每种特征选择算法对于特征的关注不同，导致选取的特征也不同，这样就会导致不同特征选择算法有自己擅长的类别特征，这也是造成 IFSC-FCBF 算法并不能保证对每个类别的 F1 值都高于其他算法的原因之一。但通过比较 P，R，F1 的平均值可以看出依然是 IFSC-FCBF 算法最优。

表 4.4 中可以看出，正如上面所述的原因 FCBF 算法能够选择出对 crude 类别更有利的特征词，所以对于 crude 类别其 F1 值要优于其他算法。但是另外四个类别 FCBF 的分类效果都不理想。money-fx 类别的特征词可以看出 IFSC-FCBF 的查准率，查全率和 F1 值得平均值比 DFS 高出 1%，比 IG 高出 2%。总的来说，在两个英文数据集上，我们提出的算法在宏 F1 值较高时，对于每个类别的分类也是最理想的。

表 4.5 中可以看出，对于复旦大学语料库，有趣的是 IFSC-FCBF 和 DFS 特征选择算法的几乎相同，IG 算法次之。表 4.6 中，IFSC-FCBF 算法对 health，education 和 tourism 三个类别的选择的特征更加有效，DFS 则对 sport 类更加有效，最后，IG 和 DFS 得到了几乎相同的平均值，但是由图 4 的宏 F1 值对比可知，此时 IG 算法要略胜 DFS。总的来说，在中文数据集上，我们的算法同样能更有效的选择特征。

由于有些时候 DFS 算法和 IFSC-FCBF 算法的性能很接近，所以我们把时间复杂度作为第二层次的比较标准，特征维度选择 300，重复十次实验取得运行时间的平均值，如图 5 所示：

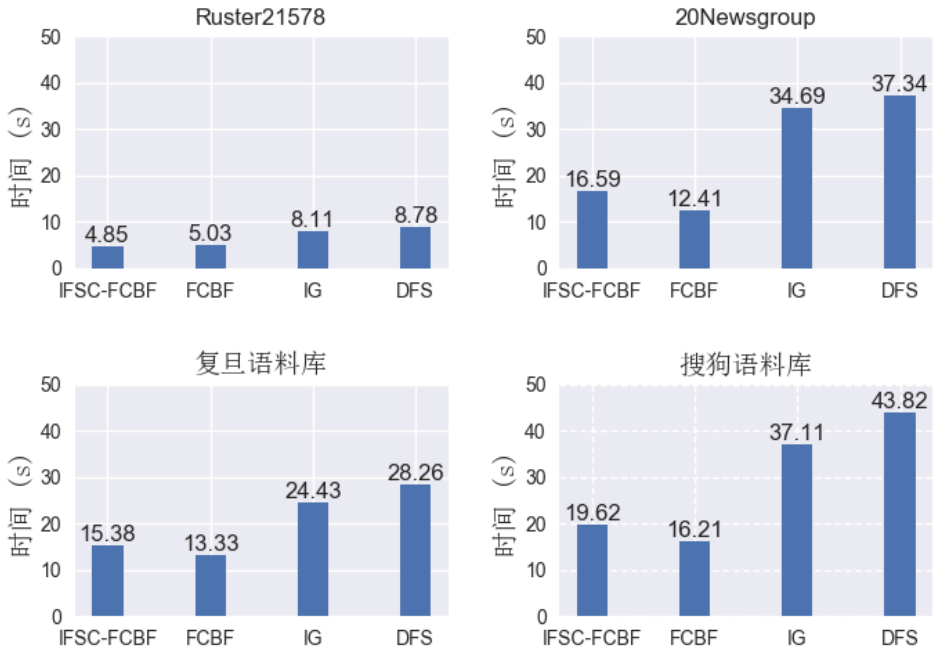

图 4.5：各算法运行时间比较

因为四个数据集所包含的特征词个数不一样，经过我们的统计 Ruster21578 包含大约 5000 个特征词，Fudan 语料库包含大约 12,000 个特征词，20newsgroup 包含大约 13,000 个特征词，Sougou 语料库包含大约 18,000 个特征词.可以看出因为每个数据集包含的特征词不数量不同，导致在每个数据集上进行特征选择所用的时间也不同，FCBF 算法的运行时间最小，因为其消除特征容易过剩所以并不能保证很高的准确率，相比之下，IFSC-FCBF 算法在运行时间上要比 IG 和 DFS 缩短一倍的时间，而 IG 和 DFS 算法的时间复杂度很接近，综合以上的分析，我们提出的 IFSC-FCBF 特征选择算法在保证运行时间比较小的情况下，同时能够选择出更加有效的特征，对文本分类任务有较大的提升。

由于第二章与第三章统一使用文档频率作为特征选择的标准，所以为了进一步探究 IFSC-FCBF 特征选择算法在朴素贝叶斯文本分类系统中的有效性，本节中将前面提出的算法结合在一起进行了一次实验。实验仍在四个语料库上进行。本次实验中比较两种算法，一种是基于 IFSC-FCBF 特征选择的 IGDC-DWNB 算法，作为对此另外一种是基于文档频率(DF)特征选择的 IGDC-DWNB 算法。两种算法的整体示意图如 4.6 所示。

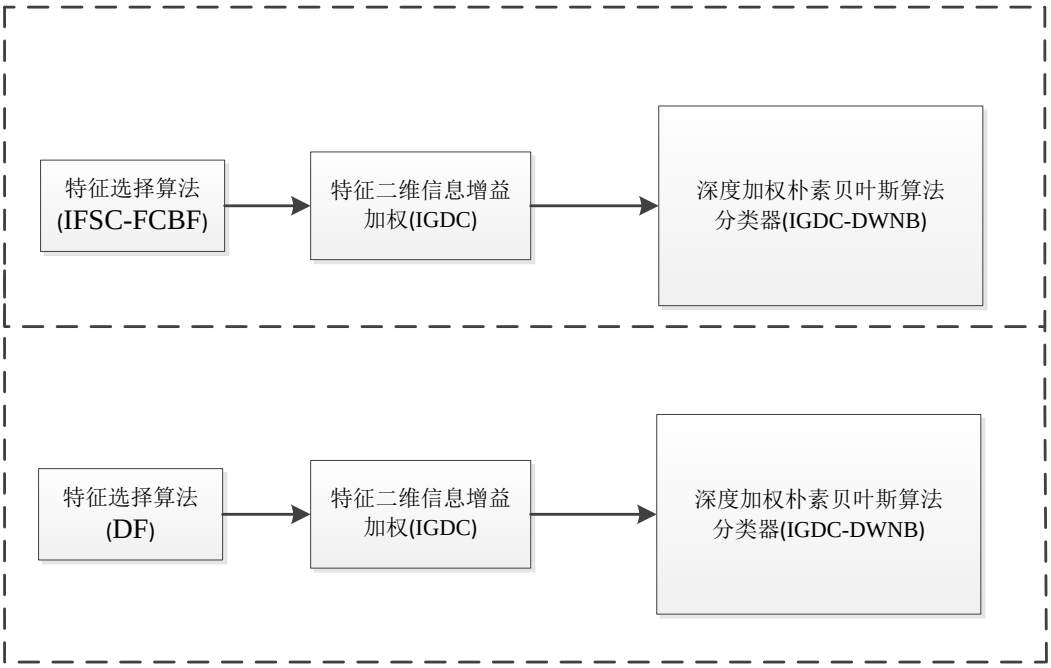

图 4.6 朴素贝叶斯文本分类整体算法示意图

在实验环境不变的情况下，两种算法在各个数据集上的宏F1值图下图4.7-4.10所示。

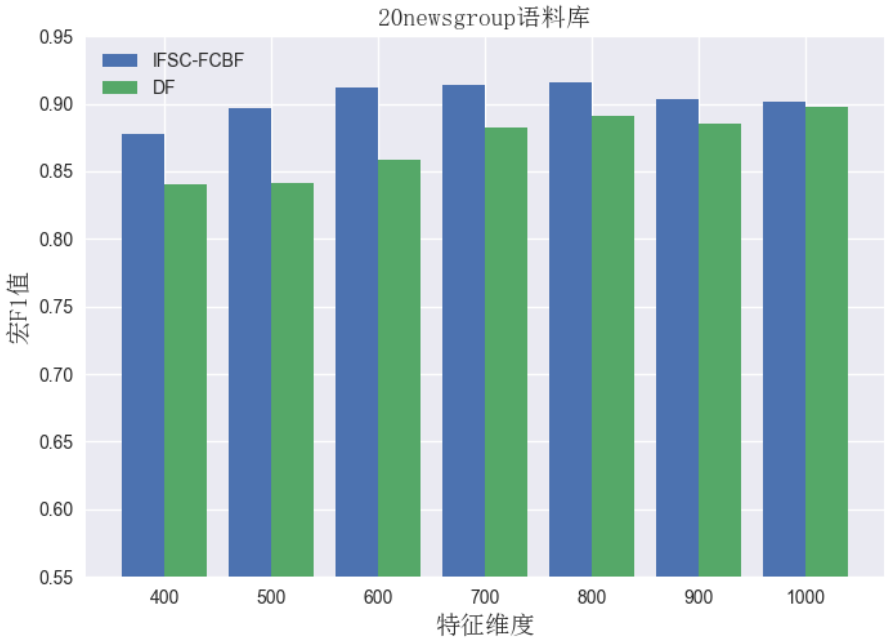

图 4.7. 20newsgroup 数据集的算法性能比较

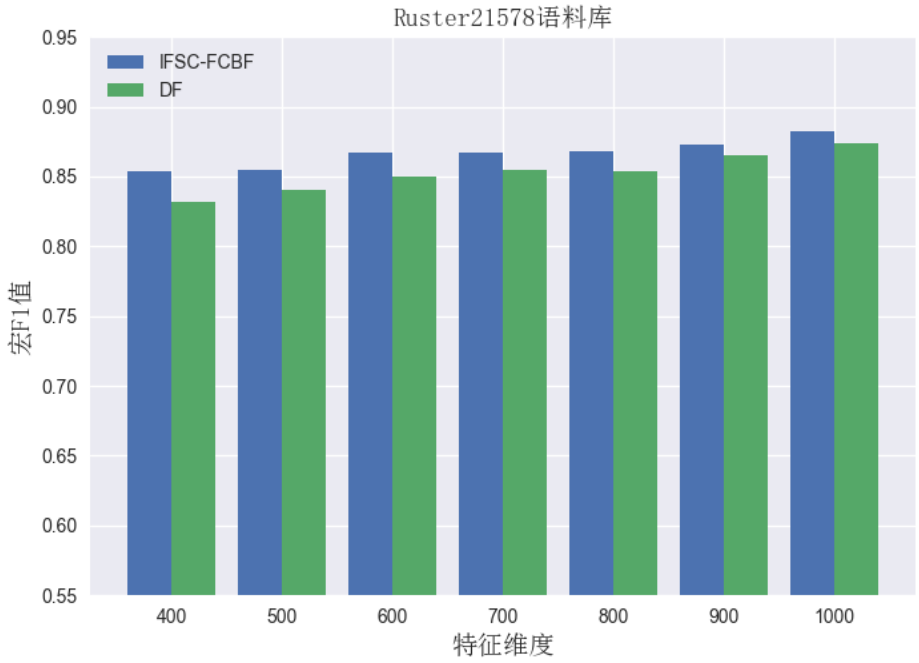

图 4.8. Ruster21678 数据集的算法性能比较

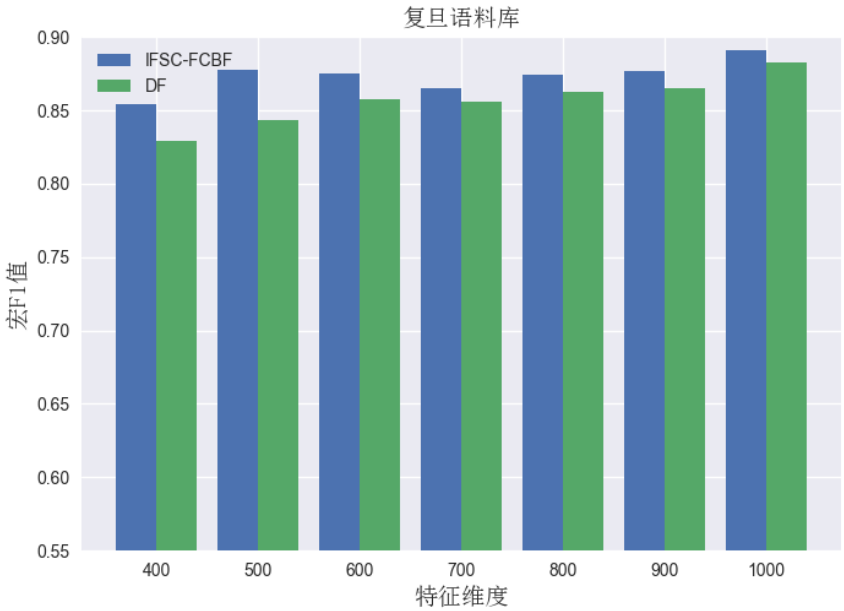

图 4.9. 复旦大学语料库的算法性能比较

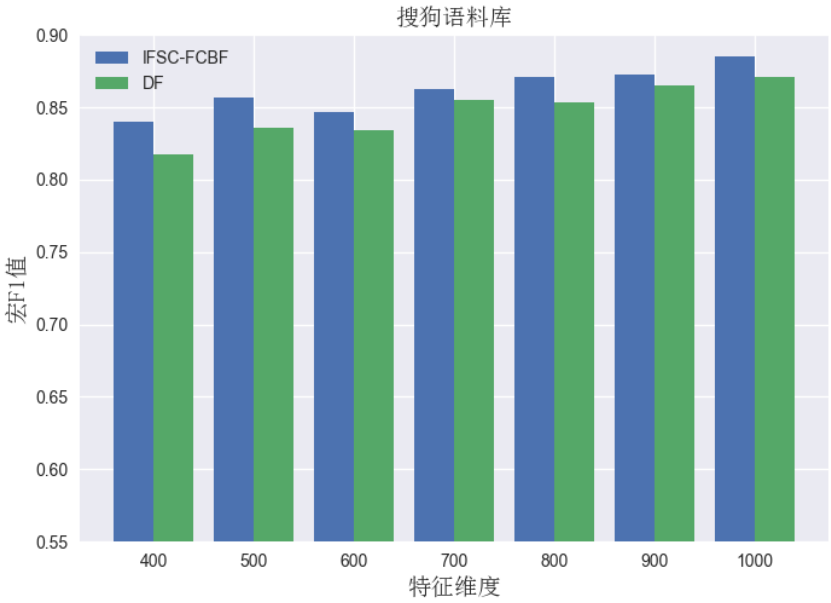

图 4. 10.　搜狗语料库的算法性能比较

由图 4.7-4.8 可以看出，在使用本章提出的 IFSC-FCBF 特征选择算法之后，分类系统的宏 F1 值在英文数据集上具有更好的效果，而两种方法随着特征维度的提升整体都是呈现增长的趋势，这是由于随着特征维度的增加，引入了更多有用的特征使得分类效果更佳。在 20newsgroup 数据集上当特征在 400-600 时提升最为明显，整体提升了 2%-3%；在 Ruster21578 数据集上的提升效果比较平稳，整体提升了 1%-2%，说明了本文提出的文本分类算法在英文数据集上的有效性。

由图 4.9-4.10 可以看出，在两种算法上随着特征维度的增加都是呈现上升的趋势，同时也能明显看出本章提出的 IFSC-FCBF 算法在整个文本分类系统中的有效性。在复旦语料库上本章提出的 IFSC-FCBF 算法比 DF 的宏 F1 值提升了 2%左右，在搜狗语料库上宏 F1 值也提升了 1%-2%，说明了本文的整个文本分类算法在中文数据集上的有效性。

两种文本分类算法对于每个类别的分类效果如下表 4.7-4.10 所示。

表 4.7 20newsgroup 数据集各类别分类效果比较　特征维度 500

	IFSC-FCBF+IGDC-DWNB			DF+IGDC-DWNB		
类别	P	R	F1	P	R	F1
alt.atheism	0.8861	0.9722	**0.9271**	0.8421	0.8888	0.8648
comp.graphics	0.8873	0.8513	**0.8689**	0.8169	0.7838	0.8000
misc.forsale	0.7831	0.8441	**0.8125**	0.6534	0.8571	0.7416
rec.autos	0.9620	0.8172	**0.8837**	0.9577	0.7312	0.8293
sci.crypt	0.9294	0.9634	**0.9461**	0.8795	0.8902	0.8848
talk.politics.guns	0.9277	0.9390	**0.9333**	0.9231	0.8780	0.9000
平均值	**0.8959**	**0.8978**	**0.8952**	0.84545	0.8381	0.8367

表 4.8 Ruster21578 数据集各类别分类效果比较　特征维度 500

	IFSC-FCBF+IGDC-DWNB			DF+IGDC-DWNB		
类别	P	R	F1	P	R	F1
Crude	0.9828	0.9828	**0.9828**	0.9828	0.9828	**0.9828**
Acq	0.8793	0.8947	**0.8869**	0.8750	0.8596	0.8672
Interest	0.6000	0.7368	**0.6614**	0.5735	0.6842	0.6239
Wheat	0.9830	1.0000	**0.9914**	0.9830	1.0000	**0.9914**
Money	0.7543	0.6232	**0.6825**	0.7167	0.6232	0.6667
Earn	0.9310	0.8852	**0.9075**	0.9152	0.8852	0.9000
平均值	**0.8550**	**0.8537**	**0.8520**	0.8410	0.8391	0.8386

表 4.9 复旦大学语料库各类别分类效果比较　特征维度 500

	IFSC-FCBF+IGDC-DWNB			DF+IGDC-DWNB		
类别	P	R	F1	P	R	F1
环境	0.9843	0.8077	0.8873	0.9701	0.8333	**0.8965**
交通	0.5431	0.9692	0.6961	0.5714	0.9230	**0.7058**
教育	0.9642	0.8804	**0.9204**	0.9277	0.8369	0.8800
军事	0.9692	0.8182	**0.8873**	0.8732	0.8051	0.8378
经济	0.9041	0.7857	**0.8408**	0.8857	0.7381	0.8051
体育	0.9871	0.9167	**0.9506**	0.8809	0.8809	0.8809
平均值	**0.8920**	**0.8629**	**0.8637**	0.8515	0.8362	0.8343

表 4.10 搜狗实验室语料库各类别分类效果比较　特征维度 500

	IFSC-FCBF+IGDC-DWNB			DF+IGDC-DWNB		
类别	P	R	F1	P	R	F1
健康	0.9677	0.7692	**0.8571**	0.8421	0.8205	0.8311
教育	0.8971	0.8000	**0.8458**	0.8468	0.7833	0.8138
军事	0.7531	0.9682	0.8472	0.8161	0.8809	**0.8473**
旅游	0.8823	0.8750	**0.8786**	0.8859	0.8416	0.8632
体育	0.9433	0.8621	**0.9009**	0.9043	0.8965	0.9004
文化	0.7443	0.8182	**0.7795**	0.7307	0.7851	0.7569
平均值	**0.8646**	**0.8487**	**0.8515**	0.8376	0.8346	0.8354

经过仿真实验可以看出在英文数据集上，即表 4.7-4.8 所示，对于不同的类别的分类效果，本文的 IFSC-FCBF 与 IGDC 深度加权朴素贝叶斯算法相结合之后具有更高的查准率，查全率和 F1 值，在 20newsgroup 数据及上平均值高出 5%到 6%。在 Ruster21578 数据集上平均查准率，查全率和 F1 值要提升 1%到 3%，显著说明了本章提出的算法又进一步提高了朴素贝叶

斯文本分类器的性能。在中文数据集上，从表 4.9-4.10 可以看出，在复旦大学语料库上虽然环境与交通两个类别的 F1 值较低，但两种算法相差都不明显，但在剩余四个类别上 IFSC-FCBF 与 IGDC 深度加权朴素贝叶斯算法却有显著的提升，在查全率，查准率和 F1 值三个指标上平均值提升了 3%到 5%。在搜狗语料库上，只要一个类别落后于 DF 并且相差很小，在三个指标上平均值也提升了 1%到 3%。综上所述，本文提出的三种改进方式都能明显的提升朴素贝叶斯文本分类算法的性能。

4.4 本章小结

本章给出了一个全新的特征选择算法以解决传统特征选择算法不能去除冗余特征的缺点。结合相关文献对快速相关性过滤算法(FCBF)进行了改进，提出了改进的自定义特征维度的快速相关性过滤算法(IFSC-FCBF)。该算法在原始 FCBF 算法的基础上对相关性计算公式进行优化，并改进了算法的执行流程能够保留自定义维度的特征数目。从实验仿真结果可以看出，IFSC-FCBF 算法在选择特征维度相同时能够使朴素贝叶斯算法更有效的对文本进行分类，自定义特征维度的方式克服了原始 FCBF 算法筛选特征过于稀少的缺点，同时保留了快速去除冗余特征的优点，同其他比较有效的特征选择算法相比，文章提出的 IFSC-FCBF 也能具有更好的效果。所以本章提出的方案能够有效解决去除冗余特征的问题，并提高朴素贝叶斯文本分类算法性能。

第五章 总结与展望

5.1 本文总结

现代互联网以及大数据时代的到来，导致互联网上的文本信息急剧增长，人为的处理大量的文本信息进行分类是不现实的，所以人们对于信息处理的能力的要求也不断提高。文本分类技术作为自然语言处理的一个分支，使用人工智能算法自动的对文本信息进行分类越来越受到人们的关注，而朴素贝叶斯分类器由于其简单性及计算的有效性一直在文本分类领域中占有很重要的地位。但是由于传统的朴素贝叶斯分类器是在特征之间相互独立的假设下才会成立，所以一定程度上影响了朴素贝叶斯实际的分类效果。本文一方面探究削弱朴素贝叶斯特征独立性假设的方法来进一步提升文本信息的分类能力，另一方面寻找一种更加有效的特征选择方法来对海量的数据进行特征提取，如果不进行特征提取就会增加分类系统的负担，降低分类器的性能，所以本文分别从文本分类系统的三个方向进行处理，提出了基于 IGDC 特征加权的朴素贝叶斯文本分类算法(IGDCNB)，基于 IGDC 深度加权的朴素贝叶斯文本分类算法(IGDC-DWNB)，改进的自定义特征维度的快速相关性过滤（IFSC-FCBF）算法。本文所研究的具体内容如下：

（1）研究并改进了朴素贝叶斯特征加权算法模型，从两个维度对特征信息增益的计算方式进行改进，提出了基于 IGDC 特征加权的朴素贝叶斯文本分类模型。该模型通过全新的方式计算特征在每个类别和每个文档中的信息增益，并通过线性归一化的方式结合了两个维度的信息，通过仿真实验表明，本文提出的方法大大削弱了朴素贝叶斯的特征条件独立性假设，具有更好的宏查准率，宏查准率以及宏 F1 值。

（2）研究了朴素贝叶斯深度加权模型，总结了已有深度加权模型以及朴素贝叶斯模型自身的缺陷，改进了朴素贝叶斯模型条件概率的计算方式，将二维信息增益 IGDC 应用于朴素贝叶斯的深度加权中，提出了基于 IGDC 深度加权的朴素贝叶斯文本分类模型。仿真实验表明基于 IGDC 深度加权的朴素贝叶斯文本分类同其他算法相比，具有更高的鲁棒性，同等环境下具有更高的宏查准率，宏查全率以及宏 F1 值，进一步削弱了其特征条件独立性假设。

（3）综述了 FCBF 算法的应用领域及其在文本分类中存在的缺陷，改进了特征相关性的计算方式，并优化了原始 FCBF 算法步骤，提出了改进的自定义特征维度的快速相关性过滤（IFSC-FCBF）的朴素贝叶斯文本分类算法。最后的仿真实验表明在保证特征维度相同时，能够更加快速的选择出更加优越的特征，并且消耗更少的时间，同时在引入了 IFSC-FCBF 特

征选择算法后，与普通的特征选择算法相比，该算法具有更高的宏查准率，宏查全率和宏 F1 值，进一步的提升了朴素贝叶斯文本分类系统的性能。

5.2 研究展望

基于朴素贝叶斯的文本分类算法，针对其特征独立性假设，本文首先从削弱假设入手，结合了特征的二维信息增益来削弱独立性假设；其次将特征的二维信息增益引入朴素贝叶斯的条件概率计算公式中，对朴素贝叶斯模型进行改进；最后从特征选择算法入手，提出了更加有效的 IFSC-FCBF 特征选择算法。实验仿真结果验证了以上三种改进都有效改进了朴素贝叶斯文本分类算法的性能。限于时间较短和个人精力有限，还遗留有一些问题需要其他研究者进一步深入研究。

（1）本文中提到的特征使用的是通用的 one-gram 模型，将来可以从 bi-gram 入手研究，或者 n-gram 入手研究。

（2）第二章中整合改进了特征两个维度的信息增益，或许有其他更有效的改进方式。

（3）本文的加权方式都是使用信息增益，还可以尝试使用其他多种方法结合来进行加权。

（4）本文针对朴素贝叶斯模型的改进是在条件概率的计算公式上，或许可以在先验概率和模型迭代方式上对朴素贝叶斯的模型进行改进。

（5）本文的特征提取方法是在 FCBF 算法的基础上进行改进，而 FCBF 在其他领域的改进方式或许也适用于文本分类领域。

参考文献

[1]. 董露露.基于特征选择及 LDA 模型的中文文本分类研究与实现[D].合肥：安徽大学,2014.

[2]. 中国互联网络信息中心（CNNIC）在京发布第 42 次《中国互联网络发展状况统计报告》

[3]. 江铭虎. 自然语言处理[M]. 高等教育出版社, 2007

[4]. ChowdhuryGG.Natural language processing[J].Annual Review of Information Science&Technology,2003,37(37):51-89.

[5]. Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2000, 5(4): 1 - 18.

[6]. Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques[J]. Biomedical Engineering Online, 2011, 5:51(1): 95–97.

[7]. Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[J]. Journal of the American Society for Information Science & Technology, 2008, 2(2): 96-102.

[8]. Manning C D, Raghavan P, Tze H. Introduction to Information Retrieval[M]. 人民邮电出版社, 2010: 824-825.

[9]. Holzinger A. Interactive Machine Learning (i ML)[J]. Informatik-Spektrum, 2016, 39(1): 64-68.

[10]. 曹玲玲，贝叶斯分类方法的对比研究与改进算法，学位论文，西安，西北大学，2011

[11]. Luhn, H. P . The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2):159-165.

[12]. M.E. Maron, J.L. Kuhn. On relevance, probabilistic indexing and information retrieval[J]. Journal of Association for Computing Machinery, 1960, 7(3): 216-244

[13]. Salton G , Wong A , Yang C S . A Vector Space Model for Automatic Indexing[J]. Communications of the Acm, 1974, 18(11):613-620.

[14]. Lodhi H , Saunders C , Shawetaylor J , et al. Text Classification using String Kernels[J]. Journal of Machine Learning Research, 2002, 2(3):419-444.

[15]. 李凯齐, 刁兴春, 曹建军. 基于信息增益的文本特征权重改进算法[J]. 计算机工程, 2011, 37(1):16-18.

[16]. Lei T, Zhang Y, Wang S I, et al. Simple Recurrent Units for Highly Parallelizable Recurrence[J]. 2018.

[17]. Zobel J , Moffat A . Inverted files for text search engines[J]. ACM Computing Surveys, 2006, 38(2):6-es.

[18]. Dong N, Eisenstein J. A Kernel Independence Test for Geographical Language Variation[J]. Computational Linguistics, 2016, 43(1).

[19]. 张伦干. 多项式朴素贝叶斯文本分类算法改进研究[D]，武汉，中国地质大学，2018

[20]. 余芳. 一个基于朴素贝叶斯方法的 web 文本分类系统:WebCAT[J]. 计算机工程与应用，2004, 40(13):195-197.

[21]. 贺鸣, 孙建军, 成颖. 基于朴素贝叶斯的文本分类研究综述[J]. 情报科学, 2016, V34(7):147-154.

[22]. 邹晓辉. 朴素贝叶斯算法在文本分类中的应用[J]. 数字技术与应用, 2017(12):132-133.

[23]. 陈治平, 王雷. 基于自学习 K 近邻的垃圾邮件过滤算法[J]. 计算机应用, 2005, 25(s1):7-8.

[24]. 孙新, 欧阳童, 严西敏,等. 基于训练集裁剪的加权 K 近邻文本分类算法[J]. 情报工程, 2016, 2(6):8-16.

[25]. 一种基于 k 最近邻的快速文本分类方法[J]. 中国科学院大学学报, 2005, 1(5):554-559.

[26]. 卢曼丽. 基于 K-means 算法的神经网络文本分类算法研究[J]. 中国管理信息化, 2014(21):80-82.

[27]. 刘钢, 胡四泉, 范植华,等. 神经网络在文本分类上的一种应用[J]. 计算机工程与应用，2003, 39(36):73-74.

[28]. 丁振国，黎靖，张卓. 一种改进的基于神经网络的文本分类算法[J]. 计算机应用研究，2008, 25(6):1639-1641.

[29]. 张世荣. 支持向量机文本分类算法研究[D]. 大连理工大学, 2007.

[30]. 郭太勇. 一种基于改进的 TF-IDF 和支持向量机的中文文本分类研究[J]. 软件, 2016, 37(12).

[31]. 刘志康. 一种改进的混合核函数支持向量机文本分类方法[J]. 工业控制计算机, 2016, 29(6):113-114.

[32].邸鹏, 段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理, 2014, 29(1):71-75.
[33].JiaweiHan, MichelineKamber. 数据挖掘:概念与技术[M]. 机械工业出版社, 2007.
[34].李忠波, 杨建华, 刘文琦. 基于数据填补和连续属性的朴素贝叶斯算法[J]. 计算机工程与应用, 2016, 52(1):133-140.
[35].D. M. Diab and K. M. El Hindi Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification[J], Applied Soft Computing, 54 (2017) 183-199.
[36].张玉芳, 陈小莉, 熊忠阳. 基于信息增益的特征词权重调整算法研究[J]. 计算机工程与应用, 2007, 43(35):159-161.
[37].李学明, 李海瑞, 薛亮,等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程, 2012, 38(8):37-40.
[38].饶丽丽, 刘雄辉, 张东站. 基于特征相关的改进加权朴素贝叶斯分类算法[J]. 厦门大学学报(自然版), 2012, 51(4):682-685.
[39].武建军, 李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法[J]. 计算机系统应用, 2017, 26(7):178-182.
[40].Zhang L, Jiang L, Li C, et al. Two feature weighting approaches for naive Bayes text classifiers[J]. Knowledge-Based Systems, 2014, 100(C):137-144.
[41].Wang S, Jiang L, Li C. A CFS-Based Feature Weighting Approach to Naive Bayes Text Classifiers[C]// International Conference on Artificial Neural Networks. Springer, Cham, 2014:555-562.
[42].Escalante H J, García-Limón M A, Morales-Reyes A, et al. Term-weighting learning via genetic programming for text classification[J]. Knowledge-Based Systems, 2015, 83(1):176-189.
[43].YANJUN LI, CONGNAN LUO, SOON M. CHUNG. WEIGHTED NAu00cfVE BAYES FOR TEXT CLASSIFICATION USING POSITIVE TERM-CLASS DEPENDENCY[J]. International Journal on Artificial Intelligence Tools, 2012, 21(01):1250008-.
[44].Zhang H, Sheng S. Learning weighted naive Bayes with accurate ranking[C]// IEEE International Conference on Data Mining. IEEE, 2005:567-570.
[45].Jiang L, Li C, Wang S, et al. Deep feature weighting for naive Bayes and its application to text classification[J]. Engineering Applications of Artificial Intelligence, 2016, 52(C):26-39.
[46].Jiang Q, Wang W, Han X, et al. Deep feature weighting in Naive Bayes for Chinese text classification[C]// International Conference on Cloud Computing and Intelligence Systems. IEEE, 2016:160-164.
[47].John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem[M]//Machine Learning Proceedings 1994. 1994: 121-129.
[48].Liu Y, Zhang J, Ma L. A fault diagnosis approach for diesel engines based on self-adaptive WVD, improved FCBF and PECOC-RVM[M]. Elsevier Science Publishers B. V. 2016 , 177 (C) :600-611
[49].Huang J, Rong P. A Hybrid Genetic Algorithm for Feature Selection Based on Mutual Information[J]. Pattern Recognition Letters, 2007, 28(13):1825-1844.
[50].Azam N, Yao J T. Comparison of term frequency and document frequency based feature selection metrics in text categorization[J]. Expert Systems with Applications, 2012, 39(5):4760-4768.
[51].Forman G. An extensive empirical study of feature selection metrics for text classification [M]. JMLR.org, 2003, 3 (2) :1289-1305
[52].Shang C, Li M, Feng S, et al. Feature selection via maximizing global information gain for text classification[J]. Knowledge-Based Systems, 2013, 54(4):298-309.
[53].Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27(8): 1226-1238.
[54].Lee C, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization[M]. Pergamon Press, Inc. 2006 , 42 (1) :155-165
[55].Uysal A K, Gunal S. A novel probabilistic feature selection method for text classification[J]. Knowledge-Based Systems, 2012, 36(6):226-235.
[56].Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]// Twentieth International Conference on International Conference on Machine Learning. AAAI Press, 2003:856-863.
[57].Liu Y, Zhang J, Ma L. A fault diagnosis approach for diesel engines based on self-adaptive WVD, improved FCBF and PECOC-RVM[M]. Elsevier Science Publishers B. V. 2016.

[58]. Chen J J, Song A, Zhang W. A Novel Hybrid Gene Selection Approach Based on ReliefF and FCBF[J]. International Journal of Digital Content Technology & Its Applications, 2011, 5(10):404-411.
[59]. Gharavian D, Bejani M, Sheikhan M. Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks [M]. Kluwer Academic Publishers, 2017, 76 (2) :2331-2352
[60]. Balakrishnan S, Narayanaswamy R. Feature selection using FCBF in type II diabetes databases [J]. Indian Journal for Medical Informatics, 2009(1).
[61]. Şen B, Peker M. Novel approaches for automated epileptic diagnosis using FCBF selection and classification algorithms [J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2014, 21(Sup.1):2092-2109.
[62]. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[M]. Pergamon Press, Inc. 1988.
[63]. 李凯齐, 刁兴春, 曹建军. 基于信息增益的文本特征权重改进算法[J]. 计算机工程, 2011, 37(1):16-18.
[64]. Croft J M P W B. A Language Modeling Approach to Information Retrieval[C]// 1998.
[65]. Kibriya A M , Frank E , Pfahringer B , et al. Multinomial Naive Bayes for Text Categorization Revisited[J]. Advances in Artificial Intelligence, 2004.
[66]. 张玉芳, 陈小莉, 熊忠阳. 基于信息增益的特征词权重调整算法研究[J]. 计算机工程与应用, 2007, 43(35):159-161.
[67]. 周志华.机器学习[M].北京：清华大学出版社，2016
[68]. Bidi N, Elberrichi Z. Feature selection for text classification using genetic algorithms[C]// International Conference on Modelling, Identification and Control. IEEE, 2017:806-810.
[69]. Jiang M, Liang Y, Feng X, et al. Text classification based on deep belief network and softmax regression[J]. Neural Computing & Applications, 2016(7):1-10.
[70]. Uysal A K. An improved global feature selection scheme for text classification[M]. Pergamon Press, Inc. 2016, 43 (C) :82-92
[71]. Conneau A, Schwenk H, Barrault L, et al. Very Deep Convolutional Networks for Text Classification[J]. 2016:1107-1116.
[72]. Senliol B, Gulgezen G, Yu L, et al. Fast Correlation Based Filter (FCBF) with a different search strategy[C]// International Symposium on Computer and Information Sciences. IEEE, 2008:1-4.
[73]. Quinlan J R. C4.5: programs for machine learning [J]. 1993, 1.

附录 1 攻读硕士学位期间撰写的论文

（1）Wei He, Yun Zhang, Shujuan Yu, Wenfeng Zhu，Deep featue weighting with a novel information gain for Naive Bayes text classification , Journal of Information Hiding and Multimedia Signal Processing(JIHMSP), Vol.10, No.1, 2019;

（2）Yun Zhang, Wei He, Shujuan Yu, Wenfeng Zhu，Improved Feature size Customized Fast Correlation-Based Filter for Naive Bayes Text Classification , Malaysian Journal of Computer Science，SCI 在审；

附录 2 攻读硕士学位期间申请的专利

（1）张昀，于舒娟，何伟，基于特征二维信息增益加权的朴素贝叶斯文本分类方法，201810019705.6，2018.1，待授权

（2）张昀，于舒娟，何伟，基于改进深度加权的朴素贝叶斯文本分类方法，201810382423.2，2018.4，待授权

附录 3 攻读硕士学位期间参加的科研项目

（1）国家自然科学基金，基于深度学习的移位 MIMO”鬼”成像方法(61871234)

致谢

在南邮三年时光匆匆过去，毕业论文的撰写也已接近尾声，在此衷心的感谢给予我帮助的老师、同学、家人和朋友！

首先，忠心感谢恩师张昀老师。张老师渊博的学识、严谨的治学态度、敏锐的洞察力、翩翩的学者风范和兢兢业业、献身科学和教育事业的崇高品格给我以深刻印象，并已成为我毕生学习的榜样。在论文的研究过程中，张老师给予我充分的鼓励和大力的支持，使我学会独立从事研究工作，在科研过程中受益匪浅。感谢张老师在学习和工作上给予的鼓励和敦促，在生活上的关心和照顾，这是我不断取得进步、直至顺利完成学业的动因。在此，向张老师致以最诚挚的谢意！

同时感谢于舒娟教授在学习和科研上的细心指导，为我的论文语言和格式进行纠正修改，在论文写作过程中给我的启发！

非常感谢我的师门朱文峰、董茜茜、金海红还有我的师兄师姐张志民，陈少威，梅可以及梁颖，感谢他们为我论文写作过程中遇到的问题进行答疑，也要感谢我的朋友王龙强、李建华、祁超等，谢谢你们在我写作过程中的陪伴与支持！

最后，感谢所有帮助过我的人，谢谢你们的支持与鼓励！