华北电力大学

硕士学位论文

基于支持向量机的中文文本分类研究

Research on Chinese Text Classification Based on Support Vector Machine

杨孟英

2017 年 3 月

万方数据

国内图书分类号：TP391.1
国际图书分类号：004

学校代码：10079
密级：公开

工学硕士学位论文

基于支持向量机的中文文本分类研究

硕士研究生：杨孟英
导　　师：胡朝举副教授

申请学位：工学硕士
学　　科：计算机科学与技术
专　　业：计算机软件与理论
所在学院：控制与计算机工程学院
答辩日期：2017 年 3 月
授予学位单位：华北电力大学

万方数据

Classified Index: TP391.1
U.D.C: 004

Thesis for the Master Degree

Research on Chinese Text Classification Based on Support Vector Machine

Candidate： Yang Mengying
Supervisor： Prof.Hu Chaoju
Academic Degree Applied for： Master of Engineering
Speciality： Computer software and theory
School： School of Control and Computer Engineering
Date of Defence： March, 2017
Degree-Conferring-Institution： North China Electric Power University

万方数据

华北电力大学硕士学位论文原创性声明

本人郑重声明：此处所提交的硕士学位论文《基于支持向量机的中文文本分类研究》，是本人在导师指导下，在华北电力大学攻读硕士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签名：　　日期：　　年　　月　　日

华北电力大学硕士学位论文使用授权书

《基于支持向量机的中文文本分类研究》系本人在华北电力大学攻读硕士学位期间在导师指导下完成的硕士学位论文。本论文的研究成果归华北电力大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解华北电力大学关于保存、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版本，同意学校将学位论文的全部或部分内容编入有关数据库进行检索，允许论文被查阅和借阅。本人授权华北电力大学，可以采用影印、缩印或扫描等复制手段保存、可以公布论文的全部或部分内容。

本学位论文属于（请在以上相应方框内打“√”）：

保密□，在　　年解密后适用本授权书

不保密□

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

万方数据

摘要

随着信息技术的快速发展，信息海量增长，如何从大量数据中获取有用信息是人们急需解决的问题。信息多数是以文本的形式出现，而中文是世界上使用人数最多的语言，所以研究中文文本分类具有重要意义。文本分类可以高效的组织和管理信息，实现快速、准确的定位信息，有效的缓解了信息混乱无序的现象。文本分类的问题是维数高、稀疏性大和特征关联度高，而支持向量机在解决这些问题上具有很大的优势，因此，支持向量机广泛应用于文本分类中。但是，支持向量机也有一些缺点，例如，样本数量增多导致分类速度变慢，参数对算法的学习性能和泛化能力影响较大。目前传统的支持向量机参数的寻优方法存在一些缺陷，比如搜索能力较弱和准确率不高等问题。

本文针对以上问题，在优化支持向量机参数方面进行了详细的研究，以达到提高文本分类的准确率和减少分类时间的效果。本文的主要研究内容如下：

首先，论文系统的概述了文本分类的研究背景及意义，海内外研究和未来的发展前景，介绍了文本分类的相关理论和关键技术，对比了文本分类中常用的算法。通过实验证明，SVM 是分类效果相对较好的算法。

然后，针对支持向量机参数选取困难的现象，本文引入了萤火虫算法，并对其进行改进，将改进后的算法来优化支持向量机参数。通过实验进行对比，验证了改进后的萤火虫算法在早期全局搜索能力增强，在后期收敛速度加快，提高了算法的性能。

其次，将改进后的萤火虫算法应用于 SVM 参数优化中，并将优化后的参数应用于训练 SVM 模型中。

最后，通过实验对比标准支持向量机和改进后萤火虫算法优化的支持向量机在文本分类中的效果。实验结果显示，改进的支持向量机模型应用在文本分类时，分类速度加快，分类的精准率明显提高，增强了支持向量机的分类性能，验证了改进算法的有效性。

关键词：文本分类；支持向量机；参数优化；萤火虫算法

万方数据

Abstract

With the rapid development of information technology, the information is in the form of massive growth. How to obtain useful information from a large number of information is an urgent problem need to be solved.The information is mainly in the form of text, and the Chinese is the most widely used language in the world, so researching on Chinese text classification is of great significance.Text classification can efficiently organize and manage information, position information fast and accurately.And it effectively solves the unordered information problems.The problem of text classification is high dimensionality, sparseness and high degree of feature association.The support vector machine(SVM) has great advantages in solving these problems, therefore, the SVM is widely used in text classification.However, there are some disadvantages of SVM, for example,when the number of samples increases the speed of the classification becomes slowly, and the parameters have great influence on the learning performance and generalization ability.The problem of traditional SVM parameters optimization methods is that, search ability is weak and the problem of accuracy is not high.

In this paper, aiming at the above problems, a detailed study was made on the optimization parameters of SVM to improve the accuracy of text classification and the classification speed.The main research contents of this paper are as follows:

First of all, the paper systematically summarized the research background and significance of text classification, the current situation at home and abroad ,the future development prospects;introduced the related theory and key technology of text classification, compared with the commonly used algorithms in text categorization. Through experiments,SVM was proved to be a relatively effective algorithm.

Secondly,aiming at the difficult problem of parameter selection of support vector machine, the firefly algorithm was introduced.And an improved firefly algorithm was proposed to optimize the SVM parameters.Through experiments, the results showed that the global search ability of improved firefly algorithm was enhanced in the early,the convergence speed became fast in the latter, the performance of the algorithm was improved .

Thirdly, the improved firefly algorithm was applied to SVM parameter optimization, and the optimized parameters were applied to training SVM model.

Finally,via the experiment,compared the result of text classification between standard SVM and the improved SVM.Experimental results showed that the improved SVM model can accelerate the classification speed and improve the classification accuracy, and enhanced the classification performance of SVM.Consequently it verified the effectiveness of the improved algorithm.

万方数据

Keywords: text classification;SVM;parameter optimization;Firefly algorithm

万方数据

目录

万方数据

万方数据

第 1 章 绪 论

1.1 选题背景及意义

随着信息技术的飞速进步，信息的增长量以指数级速度增加。信息海量增长改变了人们的生活，但也给人们带来了困扰，即如何在巨大的信息量中快速、准确地得到所需要的资源[1]。过去的方式是对信息进行手动分类，但手动分类存在着一些缺陷：耗时长，浪费了大量的人力、物力和精力；分类结果的准确率不高。单纯的依靠人工分类已经满足不了当今社会人们的需求，这给文本分类带来了新的挑战与机遇。

基于当前社会的挑战与需求，出现了自动文本分类技术，利用计算机进行文本分类[2]。能够帮助用户在大量数据中识别、筛选信息，能够快速、准确地获取相关的资源[3]。网上的信息多数以文本形式进行传递，包括新闻、书籍、学术论文、档案、消息等等，因此，文本分类成为信息处理研究领域的一个热门课题。

文本分类的基本原理是，按照文本信息的内容将其划分到某个或多个已经确定好的类别中[4]。最开始的文本分类是人工分类，必须具备丰富的知识储备，才能从事这类工作；文本信息分类是一项时间和金钱花费都很大的工作，所以传统的分类方法满足不了大量信息处理的要求[5]。自动文本分类能够处理大量信息分类问题，在自然语言的理解与处理、信息过滤、邮件处理等方面有广泛的应用潜力。

随着网络信息数据累积规模的不断增大，从大量信息中识别可用资源的数据挖掘应运而生[6]。基于机器学习的文本分类方法是数据挖掘中的重要部分，主要研究信息中不能通过分析原理得到的规律，用这些规律去分析客观事物，对未知数据进行预测。可在现实应用中满足不了该原理需要的条件，因此，许多经典的、杰出的统计学方法在现实中得到的效果并不好。

20 世纪 90 年代，Vapnik 等人在统计学理论的基础上提出了一种新的机器学习方法——支持向量机(Support Vector Machines，简称 SVM)[7]。SVM 在结构风险最小化基础之上，能够使其在测试样本数量有限的条件下，同时具备较高的分类识别准确率和杰出的推广能力。运用核函数原理，将非线性问题转化为线性问题，将算法复杂度降低，便于应用于处理分类问题。SVM 解决了“维数灾难”和过学习问题，在人脸识别、图像处理、文本分类等方面得到了广泛的应用。但是 SVM 也存在一些缺点，参数对其影响很大，而且选择合适的参数比较困难；现在问题的维度越来越大，复杂性也在不断的增加。因此，SVM 是一个值得深入研究的领域。

SVM 在文本分类中具有很大的优势，处理速度快、能够进行降维处理，将复杂

万方数据

问题简单化。最初的文本分类起源于国外，是对英文进行分类，中文分类起步较晚，知识水平和技能方面都不太成熟，而且中文的构词比英语要繁杂的多[8]。但是在全球范围内中文的使用人数居于首位，而且网络技术飞速发展，资源趋于全球统一化发展，中文资源的存储量在极速增加。因此，加强对中文文本分类的研究，提高中文文本分类的效率，发展分类技术在实际生活中的应用，对全球经济文化的发展具有重大的意义。

1.2 国内外研究现状

1.2.1 文本分类概述

文本分类的过程，即把未标记的文本按照一定的规律归类到一个或多个已经定义好的类别中，由多个领域支撑和组成，包括数据挖掘、统计学理论、机器学习等学科。

文本分类在国外的研究较早，国内研究的研究时间相对较短。1958 年，H.P.Luhn 提出了词频统计的思想，利用词对文本建立索引，并用索引与文本中的词进行匹配，奠定了词频分类技术的基础。文本分类的发展历程包括 3 个阶段：文本分类理论知识的研究、以专家知识为支撑的人工分类和基于机器学习的分类阶段。

第一阶段，理论知识的研究。上世纪 60 年代初，Maron 发表了《On Relevance Probabilistic Indexing and Information Retrial》，第一次提出了文本自动分类的概念，开启了文本自动分类的学术道路，之后又有许多专家在这一方面取得了瞩目的成绩。

第二阶段，专家知识人工分类阶段。20 世纪 80 年代，此阶段文本分类主要是以工程知识为准则，利用权威定义的类别，依靠手工得到模型进行划分[9]。但是，某些规则只使用于特定的领域，有些领域就无法进行分类，导致分类效率降低。因此，文本分类仍然不能满足社会的需求，学者们仍在继续研究。

第三阶段，在机器学习和统计学研究方法之上的文本分类阶段。20 世纪 90 年代以后，此阶段的分类利用了新型的知识，得到了强大的理论知识。文本分类模型是计算机智能创建，不依赖专家的规则，使得效率明显提高。

我国对文本分类的研究发展在 20 世纪 80 年代，相关的技术与理论都不太成熟，因而起初的学习主要借鉴海外的理论和技术，但是凭借中国人的聪明才智与不断钻研，在文本分类方面的发展突飞猛进。1981 年，侯汉清介绍了当时海外文本分类的基本情况，并对计算机在文本分类工作中的应用作了深入的分析。由于中文构词不同于英文构词，而且比较复杂，不能直接照搬国外的技术。国内的学者结合中文的特征，在英文分类的基础上，将支持向量机应用于中文文本分类工作中，建立了早

万方数据

期的中文文本分类的独立系统。

到目前为止，在中文文本分类方面，我国学者进行了深入的学习与钻研。这些研究主要分两方面，一方面是基于新兴的计算智能方面的研究；另一方面是基于统计学理论的研究，随着统计学理论的加入，基于统计知识的文本分类研究浪潮高涨起来。我国在中文文本方面取得了令人瞩目的成绩，例如百度搜索引擎、新浪网的中文垃圾邮件分类系统、北京大学的人民日报语料库、清华大学的现代汉语语料库和中科院的分词系统等等[10]。

1.2.2 SVM 概述

SVM 的应用范围很广泛，最初应用于模式识别方面，后来被应用到图像检索与识别、语音识别、人脸识别和文本分类中。文本分类中常用的算法有贝叶斯算法、K 近邻算法、神经网络和 SVM 等。SVM 是基于统计学理论原理，采用结构化最小原则，在处理小样本、非线性和高维向量模式识别问题中表现出了很大的优势。SVM 具有很好的推广能力，能够和其他算法相结合，产生新的算法，解决实际问题[11]。

SVM 是有监督的机器学习算法，可通过训练得到支持向量，即能够正确划分类别的集合。在文本分类中，通过对训练集进行测试，得到样本类别，然后依据得到的类别对测试集进行分类。

SVM 中不同的参数对算法的性能有着一定的影响，SVM 的参数包括核函数、核函数的参数以及惩罚参数，不同核函数之间的差异不大，但是惩罚因子 C 和核参数σ对算法性能的影响很大。C 的作用是控制对样本的惩罚程度，值越大对错误的惩罚越重，泛化能力就会越低。σ为函数的宽度参数，控制了函数的径向作用范围，当σ比训练样本间最小距离小很多时，则所有样本都是支持向量，都能正确分类；当σ比训练样本间最大距离大很多时，所有样本都分为一类，失去学习能力。因此，惩罚因子 C 和核参数σ的选取至关重要。

1.3 本文工作

论文对基于支持向量机的中文文本进行了全面的研究，系统的介绍了文本分类的背景与意义、国内外的研究近况；文本分类的相关技术与理论；SVM 在文本分类中的应用。虽然 SVM 在文本分类中表现出了很好的分类效果，但是 SVM 也存在一些缺陷，例如，SVM 的分类性能受参数的影响很大，而选取合适的参数往往比较困难；SVM 适用于二类分类问题，对于多分类问题表现不足。针对以上问题文本进行了研究，本文的工作如下：

（1）阐述了文本分类当前的研究背景及意义，海内外的研究现状，文本分类的一般技术与理论，SVM 在文本分类中的应用。

万方数据

（2） 对文本分类中常用的算法 SVM、朴素贝叶斯算法和 K 近邻算法进行了比较，通过研究和实验，实验结果表明 SVM 是文本分类中效果最佳的算法。

（3）提出了一种基于改进萤火虫算法的支持向量机参数优化方法。先对萤火虫算法进行改进，在萤火虫位置更新公式中加入迭代次数因子，增强了算法前期的搜索能力和后期的收敛能力。

（4）本文将改进后的萤火虫算法应用于 SVM 参数优化中，并将优化后的参数应用于训练 SVM 模型中。

（5）通过实验对算法进行了比较，实验结果验证了改进后的 SVM 算法应用于文本分类后，在分类精准度和速度方面都高于标准的 SVM，验证了提出的新算法的可行性和高效性。

1.4 论文的组织安排

本文主要研究的是基于 SVM 的中文文本分类，基于改进的 SVM 提高文本分类的速度和准确率，论本的章节组织如下：

第 1 章：绪论。着重介绍了文本分类的背景意义，阐述了海内外的研究成果，SVM 在中文文本分类中存在的问题。对论文的章节进行了详细的安排。

第 2 章：文本分类相关理论与技术。首先介绍了文本分类过程所需要的步骤，包括文本预处理、文本表示、特征提取及特征加权和分类器等。然后，比较各个步骤不同方法之间的差异和优劣。

第 3 章：文本分类方法对比研究。介绍了常见的文本分类算法，然后通过对比实验，分析了三种算法的分类性能，实验结果表示 SVM 的分类成绩最好。

第 4 章：改进的 SVM 参数优化方法。引入萤火虫算法，并对萤火虫算法进行了改进，将改进后算法的对 SVM 参数进行选取。实验结果表明改进后 SVM 的分类性能得到提高。

第 5 章：实验及结果分析。介绍了文本分类实验，改进后萤火虫算法对 SVM 的参数选取，将改进后的模型应用于文本分类，并与标准 SVM 进行对比，实验结果表明，改进的 SVM 算法有效的提高了文本分类的准确率和速度，加快了分类的时间，说明了改进算法的可行性。

第 6 章：总结与展望。对全文进行总结，提出论文的不足，并对未来工作进行展望。

万方数据

第 2 章 文本分类相关理论与技术

2.1 文本分类一般过程

文本分类，就是把未分类的文本按照一定的要求，划分到预定好的类别信息中去的过程[12]。根据现有类别的文本内容，构造一个分类模型，根据这个模型将未分类的文本信息分类。文本信息分为两类，一类是用来训练模型，称为训练集；另一类是测试数据，称为测试集。利用测试集得到的类别，来评价分类模型的准确率。文本分类的一般过程如图 2-1 所示：

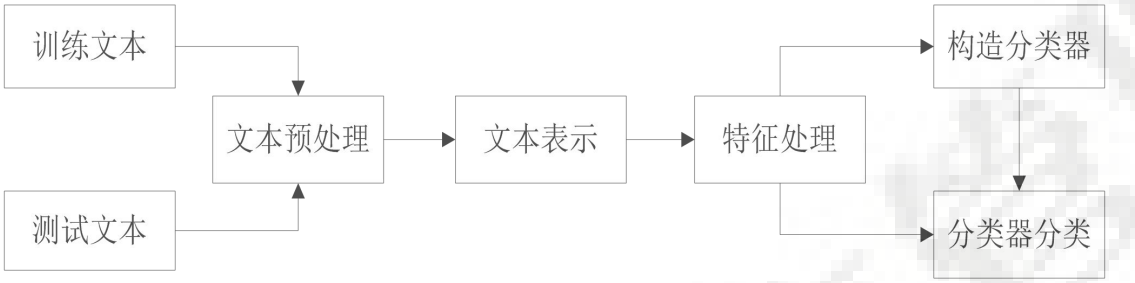

图 2-1 文本分类过程

文本分类的过程，首先将训练文本集进行预处理，主要是在信息中提取关键字，经过此步骤后，对文本分类没有影响的噪声减少；然后是文本表示，将文本信息转化为计算机可以处理的形式，计算机只能读懂 0 和 1，利用计算机系统进行文本分类，因此，要把文本转化为计算机可以读懂的形式；其次是特征处理，分为特征提取和特征加权，目的是对文本向量进行降维，为特征向量赋予不同的权值，更好的进行文本分类；最后，利用分类器对文本信息进行分类，得到测试数据集的类别。

2.2 文本预处理

文本信息中往往存在着对分类没有意义的字词、标点符号等噪声，为了提高文本分类的准确率，需要对文本进行预处理 [13]。文本信息中主要部分是文本内容，除此之外还有一些标签，比如标点符号、数字等，这些标签不存在实际意义，对分类的判断没有实际意义。

中文文本分类与英文分类的构词不同，导致了分类方法的存在差异，主要体现在文本预处理部分。两种语言最大的不同是，英文分词通过空格将两个单词分开，中文则是通过字、词语、句子和段落进行区分，没有明显的形式区分。所以中文的预处理过程比较复杂，文本预处理的步骤一般包括处理文本标记、中文分词和过滤停用词等。

万方数据

2.2.1 处理文本标记

文本数据信息不仅包括纯文本内容，还含有一些没有实际意义的标记，例如网页标签、标点符号等，或者是声音、视频等媒体。这些标记在文本分类中没有实际意义，不仅不能作为文本分类的依据，还会影响分类效率，增加分类时间，浪费分类资源。例如，网络信息中的视频，有些只是为了吸引浏览者，但是与文本信息的主题类别无关，这样的视频就干扰了信息分类。

为了减轻文本分类的负担，在预处理部分需要处理文本标记，将这些标记过滤掉，现在网络信息越来越多，文本信息大多数以网页的形式呈现。因此，删除的部分主要是网页的代码标签、脚本信息和视频声音等媒体信息。文本标记处理后，文本数据的噪声相对减少，文本信息的质量提高，为之后的文本分类提供了有利的条件。

2.2.2 中文分词

中文构词与英文不同，英文内容中每个单词间有明显的空格，利用空格将单词分开，进行分词处理就会比较简单。而中文文本字词之间没有明显的分割界限，都是连续的字词，只有句子和段落间有标点和空格，所以中文分词比较复杂。中文分词时会通过选择字、词或词组作为特征项，通过实验，普遍认为以词作为特征项分类效果高于字和词。词组表达的含义明确，而且不容易出现同义的情况，但是文本信息中词组出现的频率较低，导致信息量减少，影响分类准确率[14]。因此，要从文本信息中提取关键信息，必须对文本进行分词处理，在中文文本分类步骤中，中文分词是一个较为重要的技术。

根据不同的研究方向，常见的分词方法有以下三类：

（1）基于概率统计的算法

基于概率统计的分词思想是：在文本信息中，统计相邻字出现的次数，把次数作为评价相邻字为一个词的评价标准。对文本中相邻字组合的频率进行统计，计算互现信息，互现信息代表了字与字之间的关联程度。对表示关联程度的次数设置一个阈值，当出现的次数高于阈值时，表示相邻字同时出现的频率很高，可以构成一个新的词，能够体现文本的特征，可以作为分类的依据。

基于概率统计的分词具有很好的实用性，不需要建立语料库，也不需要以专业知识为基础，根据上下文信息就可以分词。但这种方法也有一定的局限性，出现频率高的相邻字不一定是常用的组合，不一定有实际的意义，不仅不能为分类提供依据，还影响分类效率；而且可能会导致特征项数量多，空间维数太高，文本分类效率低下等问题。

万方数据

（2）基于词典匹配的算法

有些常用词以及固定的词都收集到词典中，基于词典匹配的分词算法就是根据词典原理，将大量的词收集在一起成为词典。算法的基本思想是根据一定的规则将信息中的字符串和词典中的词进行匹配，若匹配成功则将该其提取出来，成为一个词[15]。

基于词典匹配进行分词的方法思想简单、分词的效率较高，但是完全依赖词库，分词比较机械，而且现在词库信息不完整、规则也没有完全统一，利用此方法进行分词有些困难，所以这种方法很难处理大量文本分词任务。

（3）基于理解的分词算法

基于理解的分词方法是建立在知识语言基础之上的，主要是让计算机模拟人对语言的理解，按照人类的方式分词，是一个知识推理的过程，需要对语句、语法进行分析。因此，建立在大量语言基础之上，通过文本信息的上下文内容对分词算法进行指导，并且处理歧义分析。

基于理解的分词方法建立在语言理解之上，由于汉语语言的复杂多变，结构多样化，将自然语言转化为机器能够直接理解的形式是一项不小的挑战。

2.2.3 过滤停用词

中文文本信息中总会有一些没有实际意义的词，这些词统称为虚词。虚词在文本信息中出现的概率往往较大，却不能为文本分类提供帮助，还会增加文本分类的复杂度，将它们称为停用词。

停用词可以分为两类，一类是在文本中高频率出现的词，几乎包含在所有文本中，没有区分能力，这样的词对分类没有任何帮助作用。另一类是代词、副词、语气助词等虚词，如“呀”、“他”和“是”等对分类没有意义的词。

将停用词删除能够增加存储空间，减少一定的噪声和降低空间向量的维数。过滤停用词的一般做法是建立停用词表，将分词得到的关键字与停用词表相匹配，匹配成功则表示其是停用词，则从关键字中将其去掉；若匹配失败，表明不是停用词，则在关键字中保留[16]。

2.3 文本表示

文本信息的结构很复杂，计算机是无法直接对文本信息进行处理的，所以要将文本信息转换成计算机能够读懂的形式。文本表示是文本分类的前提和基础，在文本分类中具有很大意义。目前常见的模型有布尔模型、向量空间模型、概率模拟模型，下面对三种经典模型进行详细介绍：

（1）布尔模型

万方数据

布尔模型产生较早，原理比较简单，是建立在集合论和布尔代数原理之上的文本表示模型。是一种以特征项严格匹配的模型，特征项的权重只有“0”和“1”两种情况。布尔模型中，特征项在文档中只有两种状态，即出现或者不出现，相应的变量的取值集合只有“true”和“false”。 文本信息通过以上方式表示，转化为“0”和“1”的集合，特征项出现在文档中用“1”表示，不出现在文档中用“0”表示。

布尔模型在上世纪 60 年代有了很大的发展，出现了许多商用系统。布尔模型查询功能由特征项和 “与”、“或”、“非”等逻辑运算组成，该模型的优点是原理简单，结构清晰，便于表达和运算。但是布尔模型也存在着一些缺陷，在进行检索时，采取的是精确匹配，只有相关和不相关两种情况，一些模糊匹配和相关性的特征项被排除，造成了特征项的丢失，限制了文档的检索功能；文本信息含量较大时，很难将文档表示为布尔表达式，实用性能有待提高。

（2）向量空间模型

向量空间模型（Vector Space Modal，VSM），在 20 世纪 60 年代由 Salton 等人提出。提出之后被广泛应用于信息检索和数据挖掘等范围，是最简便、高效的文本表示模型之一。

在向量空间模型中，文本集合 D 由一组文本（$d_1,d_2,\ldots,d_n$）组成，其中 d_i 表示文档集中的某一文档；每个文本由多个特征项（$t_1,t_2,\cdots,t_n$）组成，t_i 表示文本信息中的一个特征项。特征项的权值由 w_i 表示，则文档可表示为 $d=d\{w_1,w_2,\ldots,w_n\}$，$w_i$ 表示特征项 t_i 的的权重，$1\leq i\leq n$。文本分类问题转化成若干向量间的计算问题，降低了文本分类的复杂性。文本信息间的相似度可以通过向量间的距离来表示，通常使用的方法有向量内积法、Jaccard 系数法和余弦夹角系数法。

常用的方法是余弦夹角法，就是计算两个向量间夹角的余弦值，用余弦值表示相关度，余弦值越大，说明文本间的相似度越大。计算公式如下：

$$sim(D_1, D_2) = \sum_{i=1}^{n} w_{1k} * w_{2k} \tag{2-1}$$

$$sim(D_1D_2) = \frac{w_i * w_q}{\sqrt{\sum_{j=1}^{n} w_{ik}^{\ 2}}\sqrt{\sum_{j=1}^{n} w_{jk}^{\ 2}}} \tag{2-2}$$

上式中，w_{ij} 表示第 i 个文档的第 j 个特征的权重。

向量空间模型将复杂的文本信息匹配问题转为向量间的计算问题，将问题简单化，减小了计算量，具有很强的实用性。文本内容以向量的形式转化为数值形式，能够利用更多的数学方法计算向量间的关系，大大提高了自然语言的处理能力。知识表示一直是知识处理中的重要难题，特别是研究对象是自然语言的处理，在知识分析和获取方面一直以来都是困扰人们的难题。

万方数据

向量空间模型的优点有：结构简单、易于操作计算；不涉及语义、语法等语言复杂问题，可以忽略句子间的关系，段落间的联系，使自然语言的处理变得简单；特征词权重的加入，提高了分类的真实性，使文本分类的检索性能得到明显提高；部分匹配方法更符合自然语言的语法特征，筛选后的文本集更符合用户的要求[17]。

（3）概率模型

概率模型也是常用的文本表示模型之一，相比于前两种模型，它是基于统计学和概率原理，由 Robertson 和 SparckJones 提出。文本集合中特征项间没有任何有关的联系，都是独立的属性，而且与其在文本中的顺序没有关系。概率模型基于以上特征，将用户的兴趣与文本集中文本出现的概率进行统计，并将统计结果按从大到小顺序进行排列，根据排列的先后顺序赋予特征项不同的权值，然后进行文本分类。

概率模型的基本思路：文本集合 D=D{$d_1,d_2,\ldots,d_m$}，d_i 表示文本集合中的一个文档，类别 C=C{$c_1,c_2,\ldots,c_n$}，c_j 表示一个类别；查询字符串用向量表示，q=（$w_{q1},w_{q2},\ldots,w_{qm}$）；计算文档中类别的概率 P（C/D,q），其中，P 代表文本 D 与用户查询 q 的相关概率，相关和不相关的概率总和是 1。某一类别使得概率 P 最大，则文档被归类为此类别。

在信息处理过程中，概率 P 的运算转化为 P（C|t,qk），此时与分类没有关系的特征项被过滤掉。将文本信息对照计算的相关概率进行排序，相当于文本按照特征向量进行的排序。文本的概率相关性计算公式如下：

$$P(C|D,q)=\sum_i d_i \times \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \tag{2-3}$$

上式中，p_i=P(ti=1|C,q)，q_i=P(t_i=1|C,q)。

公式中，p_i，q_i 参数通过统计数量得到，公式如下：

$$p_i=\frac{c_i}{c}\ ,q_i=\frac{n_i-c_i}{n-c} \tag{2-4}$$

上式中，n 表示需要比较文本集合中包含的数量，c 为与查询有关的文本数量，n_i 表示特征项在文档中的个数，c_i 为特征项与查询相关的文本个数。通过计算文本间相关概率，对排列顺序进行调整，成为一个较好的概率排列。

概率模型没有考虑文本语言中语义、语法等复杂的关系，将文本间的相关性转化为统计后的概率信息，通过用户查询文本与匹配文本的相关概率进行分类，将复杂问题简单化。但是概率模型也存在一些缺点，比如，只是根据用户的兴趣进行统计，将概率按序排列，并没有考虑文本信息本身特征项的出现频率，可能会使信息的遗漏，影响分类的准确率。该模型对查询信息的依赖性很严重，如果查询信息发生改变，分类结果也会随着信息的改变而变化，影响分类效果，造成文本信息的分类不准确。

万方数据

2.4 特征处理

文本训练集在文本预处理之后，得到关键字的集合，称为特征集；再进行中文分词，获得更简练的特征项集合。虽然经过多次处理，但是文本中特征项的集合中的内容仍然很多，造成空间的维数很高。特征处理的过程，就是对特征集中特征进行提取，去掉弱关联词，提取强关联词构词新的特征集，再通过权重法对新的特征项赋予不同的权值，表示在文本分类中不同的贡献作用。

2.4.1 特征提取

在文本分类中，特征集包含大量的信息，对特征项进行提取和文本表示后，文本向量的空间维数变的较高，给分类带来了不小的问题，使得计算的开销很大，导致分类准确率不高。因此，要对特征集进行降维处理，维数变小后，计算量大大减小，使得文本分类的效率和准确率得到提高。

特征项具备的特点，能够准确表示文本的内容；能够和其他的特征项进行区分，有自身的特点；在文本中的数量相对较少。特征提取的过程即是对特征向量降维操作的过程。目前常见的特征提取方法有频率统计、信息增益、互信息和χ^2统计法[18]。

（1）频率统计法

频率统计包括对词频和文档频率的统计。词频统计（Term Frequency，简称 TF）指的是对一定长度的词语在文本中出现次数的统计，对分类结果进行归纳。一个词语在文本中出现的频率与文本的类别有很大关系，出现的频率越高与文本类别关系越密切。

文档频率（Document Frequency， DF）的思想是，某一特征出现的文档数占总的文档集合的数量的比例，比例越大表明出现的次数越多，特征词对文本分类的作用越大，越能够作为分类的重要依据。将这些词过滤之后，文本向量的维数降低，但是对分类的准确率没有影响。提取特征词后，向量空间的维数降低，计算量减小，因此，文本分类的时间和准确率会相应的得到提高。

文档频率统计方法原理比较简单，计算的复杂度随文本集合规模的增大而增加，比较适用于数据量较大的训练集合。但是文档频率统计方法也存在一些问题，根据特征词出现的频率将其删除，可能会失去重要的分类信息。因为，虽然有些词在文档中出现的频率低，但是往往具有较高的分类价值，对文本分类起着至关重要的作用，将这些词删除会导致分类精度下降[19]。因此，应该根据具体情况具体分析，选择合适的提取方法。

（2）信息增益

信息增益（Information Gain）广泛应用于机器学习中，表示某一特征项在文本

万方数据

中作用，出现前与出现后对文本分类的影响，定义为其出现前后的信息熵的差。

信息增益的计算方法如下：

$$IG(w) = -\sum_{i=1}^{m} P(C_i)\log P(C_i) + P(W)\sum_{i=1}^{m} P(C_i|W)\log P(C_i|W) + P(\overline{W})\sum_{i=1}^{m} P(C_i|\overline{W}) \quad (2\text{-}5)$$

上式中， P（C_i|W）表示文档包含特征项 W 时属于 C_i 类的概率，P（$\overline{W}$）表示训练集中不包含特征项 W 的文档的概率，P（$C_i|\overline{W}$）表示训练集中不包含特征项 W 时属于 C 类的条件概率，m 表示文本的类别数。

计算训练集中每个特征项的增益值，设定一个阈值，将低于阈值的特征项删除，高于阈值的特征项保留下来作为新的特征项。信息增益也存在一些缺点，当某个特征项仅出现在一个类别中时，虽然均匀分布在类中，但是增益值也会很小，这样就会影响分类的准确度。

（3）互信息法

互信息法在统计语言模型中的应用很广泛，表现的是词条与类别的关系。基本思想是，词条在某一类别中出现的频率高，在其他类别中出现的频率低，则特征与类别的互信息值越高，这个词条成为类别标签的概率越大。互信息的计算公式如下：

$$MI(t,c) = \log\frac{p(c|t)}{p(t)\times p(c)} \quad (2\text{-}6)$$

上式中，p（c|t）表示包含特征 t 同时属于类别标签 c 的文本的概率，p（t）表示训练语料中包含特征 t 的文本概率，p（c）表示训练语料中属于标签 c 的文本概率。

公式（2-7）可以近似表示为如下的公式：

$$MI(t,c) = \log\frac{A\times N}{(A+C)\times(A+B)} \quad (2\text{-}7)$$

其中，A 表示含有特征 t，属于标签 c 的文本数量， B 代表含有特征 t，不属于标签 c 的文档数量，C 代表不含有特征 t 却属于标签 c 的文档频数，N 为语料中的文档总数。

互信息法是基于特征与类别的关系作为分类依据，特征项在不同类别中出现的情况，符合分类的逻辑。

（4）χ^2 统计

χ^2 统计法经常应用于统计学。假定特征和类别符合χ^2 分布，度量特征和类别之间的相关程度。某一特征对于某一类别的χ^2 值越大，表明与类别的相关性越大，携

万方数据

带的分类信息也越多，对于分类的贡献越大。特征与类别 x^2 值的计算公式如下：

$$\chi^2(t,c)=\frac{N(AD-BC)}{(A+C)(B+D)(A+B)(C+D)} \tag{2-8}$$

其中，A、B、C 表示的含义与互信息中的含义一样，D 表示不属于类别 c 也不包含特征 t 的文档数，N=A+B+C+D。如果 AD<BC，表示类别和特征词之间是相互独立的，相关程度很小。

应用于多分类问题时，要计算特征项和每个类别之间的χ^2 值，然后再计算特征项和训练集的χ^2 值，计算方法如下：

$$\chi^2_{\max}(t)=\max_{1\le i\le m}\left\{\chi^2(t,c_i)\right\} \tag{2-9}$$

上式中，m 表示类别数目，设定一个阈值，当统计的χ^2 值低于阈值时，将其删除，高于阈值时保留下来，得到新的特征集。

特征提取方法不同，分类效果也会不同，要根据文本集的情况和提取方法的优缺点，选择最优的提取方法，达到分类效果最优的情况。

2.4.2 特征加权

文本集合中每个特征词在文本中表示的重要程度不同，将特征提取过程中提取的特征项进行赋值，根据代表的重要程度给特征项赋于权重。特征项的权值反映了对文本分类的贡献度，对类别的区分能力[20]。准确的计算特征项的权重是保证文本分类精度的前提和基础。

特征项的重要特征是完全性和区分性，特征项权重的赋值必须满足以上要求，计算的方法通常有两种，一种和出现在文档的频率成正比，一种是和出现在文档的频率成反比。常见的方法有布尔权重法、TF-IDF 权重法、基于熵的权重法和词频权重法，着重介绍布尔权重法和 TF-IDF 权重法。

（1）布尔权重法

布尔权重法的表示方法比较简单，特征项在文本中出现，权值为 1，反之权值为 0。

$$w_{ik}=\begin{cases}1 & df_{ik}>0\\ 0 & df_{ik}=0\end{cases} \tag{2-10}$$

上式中，w_{ik} 表示第 i 个文档中的第 k 个特征项，df_{ik} 表示特征项在文本中出现的次数。布尔权重法只是根据特征项的出现与否赋予权重，没有考虑特征项对于分类的区分程度，分类效果并不理想，此方法适用于特征项较少的情况。

（2）TF-IDF 权重法

在信息处理领域的应用最为广泛。主要思想是统计词语在文档中的出现频率，

万方数据

频率越高，而且词语中文本数的词越少，则表明对分类的贡献越大。

特征项频率（Term Frequency，简称 TF）的统计，指统计特征项出现在文本中的次数，文档的类别不同特征项出现的频率也不一样，出现的频率越高，特征项对文本分类的意义越大。

倒排文档频率（Inverse Documentation Frequency，简称 IDF）的主要思想是，一个在很多文档中出现的特征项对文本分类的意义，要大于在少数文档中出现的特征项。此方法减弱了高频词的作用，增加了低频词的重要性。IDF 的值越大，则表明特征项出现的越集中，对文本的区分度作用越大，计算方法如下：

$$idf(T_k) = \log(\frac{N}{n_k} + 0.01) \tag{2-11}$$

由以上 3 个因素得到 TF-IDF 权重计算方法的公式如下：

$$w(t,d) = \frac{tf(t,d) \times \log(\frac{N}{n_k} + 0.01)}{\sqrt{\sum_{t \in d} \left[f(t,d) \times \log(\frac{N}{n_k} + 0.01) \right]^2}} \tag{2-12}$$

其中，w（t,d）表示特征项的权值，tf（t,d）表示特征项在文本中出现的频率，N 为总的文本数，分母为归一化处理。

2.5 分类性能评价标准

文本分类的评价标准主要基于计算的复杂度和有效性度等方面。计算的复杂度包括时间复杂度和空间复杂度，包括分类速度和内存使用率；有效性指的是正确分类的能力，代表了分类器的分类性能，是最重要的评价标准，也是评价的主要内容。

文本分类的分类情况分为 4 类，对于特定类别 c_i，一类是属于类别 c 也被正确分类到了类别 c_i 中的文档数，用 TP 表示；一类是实际属于类别 c_i 却被分到其他错误类别的文档数，用 FP 表示；一类是属于类别 c_i 的文档被错误的分到其他类别中，用 FN 表示；一类是不属于类别 c_i 的文档也没有分类到 c_i 中的文档数。分类情况如下表所示：

表 2-1　分类统计表

	属于类别 c_i 的的文本数	不属于类别 c_i 的文本数
属于类别 c_i	TP	FP
不属于类别 c_i	FN	TN

其中，有效性的评价标准有：查准率（precision）、查全率（recall）和 F 测量（F-measure）。详细介绍如下：

万方数据

（1）查准率（P）

查准率指的是对文档进行分类后，正确分类到相应类别的文档占所有分类到某一类别文档的比例。计算公式如下：

$$P = \frac{TP}{TP + FP} \times 100\% \tag{2-13}$$

上面公式中，分子为本身类别属于类别 ci 并且被正确分类到类别 ci 中的文档数量；分母是被分到类别 ci 中的总数，包括正确分类和错误分类的文档数。

（2）查全率（R）

查全率指的是正确分类的文档数与预测类别的文档数的比例，即与本身属于类别 c_i 的文档数之比。计算公式如下：

$$R = \frac{TP}{TP + FN} \times 100\% \tag{2-14}$$

其中，分子是被正确分类到类别 c_i 中的文档数，分子为实际属于类别 c_i 的文档总数。

（3）F 测量

由于查准率和查全率是相互对立的测量标准，当查准率增加时，查全率会随之减小，两者不相容。F 测量是一种可以综合两者的测量方法，其中计算公式如下：

$$F_{\beta}(P,R) = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2-15}$$

其中，β为调整参数，通过调节参数，达到平衡和限制查准率和查全率的作用。当β取值为 1 时，F_β为 F_1 值，为查全率和查准率的平均数，公式如下：

$$F_1 = \frac{2PR}{P + R} \tag{2-16}$$

（4）宏平均和微平均

上面三种方法均针对的是单一类别分类结果的统计，若对某一分类方法进行评价时，需要综合分类结果求平均值，常用的方法有微平均和宏平均。

宏平均，计算每一类中查全率和查准率，然后求得平均值，获得整体值，此方法更多的侧重于特殊类别。微平均，建立在整体基础上，认为文档的权重相同，得到的是整体的查准率和查全率。

当训练集中类别差别较大时，两者的取值也有较高的不同；当某一类别出现频率较低时，宏平均表现了更好的优越性。

2.6 本章小结

本章首先系统的介绍了文本分类的一般过程及相关技术，然后对每个步骤进行了详细的介绍，包括文本预处理、文本表示、特征处理和分类方法等，讲解了每个

万方数据

步骤的重要方法和主要思想，最后介绍了文本分类的评价标准。为下一步的研究工作铺下了很好的基础。

万方数据

第 3 章 文本分类方法对比研究

分类算法的主要功能是训练文本分类器，属于文本分类过程中的核心部分。常见的分类算法有三类，一类是基于概率的算法，如朴素贝叶斯算法，最大熵算法等；另一类是基于 TF-IDF 权值计算方法的分类算法，这类算法主要包括 Rocchio、TF/IDF 和 K 近邻等；第三类是基于知识学习的分类算法，如决策树、人工神经网络和支持向量机等算法。根据不同情况选择不同的分类算法，下面详细分析几种分类效果较好的算法。

3.1 朴素贝叶斯算法

贝叶斯方法是建立在统计学理论之上的，在机器学习和人工智能领域应用广泛。假设属性间是相互独立的，从而预测属性之间的关系，判断文本属于某一类别的概率，根据预测结果将该文本归类到概率最高的类别中[21]。在文本分类中，假定每个属性间是相互独立的，朴素贝叶斯分类器的原理如图 3-1 所示。

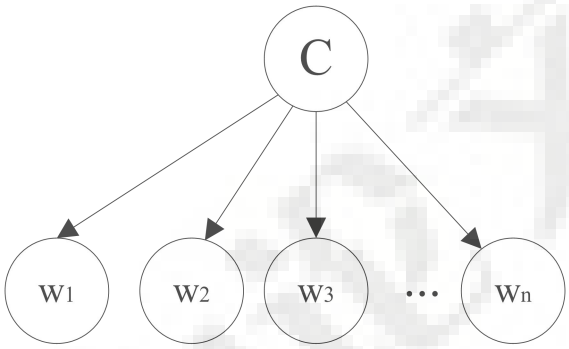

图 3-1 朴素贝叶斯分类器

朴素贝叶斯的原理是，根据先验概率、条件概率和全概率判断对象所属类别的概率，得到的概率值最大的就是对象的类别。在文本分类中，一个文档的表示为 $d_i=\{t_1,t_2,\ldots,t_n\}$，训练文本集 $D=\{c_1,c_2,\ldots,c_k\}$，根据贝叶斯公式，文本 d_i 属于类别 c_j 的概率为：

$$P(\frac{c_j}{d_i})=\frac{P(c_j)P(\frac{d_i}{c_j})}{P(d_i)} \tag{3-1}$$

$$P(d_i)=\sum_{j=1}^{k}P(c_j)P(\frac{d_i}{c_j}) \tag{3-2}$$

其中，p（cj）表示类别的先验概率，利用拉普拉斯概率原理估算得到，计算公

万方数据

式如下：

$$P(c_j)=\frac{1+Dc_j}{k+Dc} \tag{3-3}$$

其中，D_{cj} 表示 cj 中文本的数量，D_c 表示训练集中文本总数量，计算公式如下：

$$P(\frac{d_i}{c_j})=\prod_1^n P(\frac{t_I}{c_j}) \tag{3- 4}$$

文档 d_i 的类别通过公式（2-13）的结果来判定，概率值最大的即为 d_i 的最终类别。贝叶斯分类器具有的特征：

（1）算法原理简单，比较容易实现。

（2）分类器健壮性好，对干扰数据敏感，在计算过程中，这些数据会均分其他数据。

（3）分类器要求属性相互间是独立的，在实际生活中，属性完全独立的数很少，标准数据不易满足。

（4）计算过程中进行的概率估计不易实现，概率修正用到的算法比较复杂，计算较困难。

3.2 k 近邻算法

K 近邻算法（K-Nearest Neighbor，简称 KNN），建立在统计学理论之上，原理比较简单，分类效果也较好[22][27]。算法的基本思路是：把每一个样本看作一个点，求测试样本与训练样本的距离时，简化为点与点间的距离问题，选取最近的样本点为测试样本点的标签。

用向量表示文本信息，样本点间距离转化为向量间距离。常用的方法为度量向量间的欧式距离，公式如下：

$$D(d_1,d_2)=\sqrt{(w_{11}-w_{21})^2+\cdots+(w_{1n}-w_{2n})^2} \tag{3-5}$$

其中，w_{1i} 表示第一个文档的第 i 个特征项的权值，w_{2i} 表示第 2 个文档的第 i 个特征项的权值。

文本相似度的公式如下：

$$sim(d_1,d_2)=\frac{\sum_{k=1}^{n} w_{1k}\times w_{2k}}{\sqrt{\sum_{k=1}^{n} {w_{1k}}^2}\sqrt{\sum_{k=1}^{n} {w_{2k}}^2}} \tag{3-6}$$

其中，d_1 是测试数据的特征向量，d_2 是第 2 类的中心向量；n 表示总的特征维数；k 表示当前向量的维数，k 的值一般先设定初值，然后在测试中确定实验值。

万方数据

K 近邻算法原理简单，没有参数易于实现，在基于统计的模型中表现的效果很好。由于直接将文本与文本进行比较，减少了因为特征项选择不当造成的分类不准确问题。在文本分类中，K 近邻也存在一些缺点，对 k 值的依赖很大，选取不当会对分类结果造成不小的影响；相似度的计算使得空间维数较高，计算量偏大，需要的时间增多。

3.3 支持向量机算法

支持向量机是由 Vapnik 等人在 1995 年提出，基于统计学理论的 VC 维理论和结构风险最小化原理，是基于两类问题之上的机器学习方法。两类问题分类时，支持向量机将构造一个最优分类面，能够使分类间隔最大，将两类样本分割开[10][36]。

支持向量机在解决小样本、高维向量空间时有很大优势，有效地解决了维数灾难问题。其基本思想是先找到一个分类面，然后将两类样本分割开。当线性可分时，最优分类面不仅要求将两类问题正确分割开，还要求分类间隔最大；线性不可分时，一个超平面不能把两类样本完全分开，可以引入松弛变量，此时得到的最优分类面为广义最优分类超平面。

3.3.1 线性可分支持向量

当线性可分时，最优分类面不仅要求将两类问题正确分割开，还要求分类间隔最大。如图 3-2 所示，实心点与空心点分别代表两种样本，中间的虚线为分类线，也叫最优超平面，H1 和 H2 上面的点即为支持向量（Support Vector）[13]。

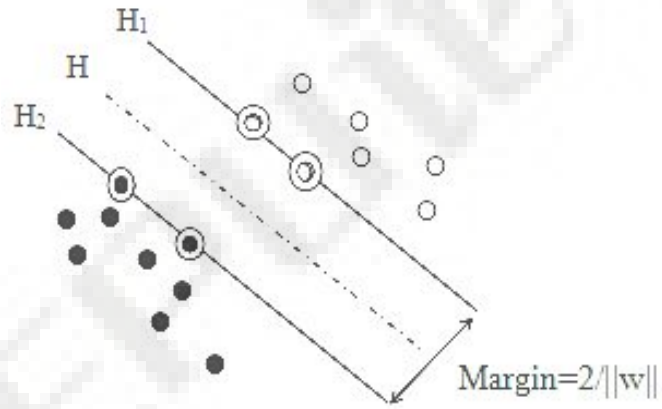

图 3-2 支持向量机原理

上图中，H 是能够将两类正确分开的分类面，H 面只是一个统称，是广泛存在的不是仅有的，对分类面 H 向上或者向下平移，与样本点刚刚接触的分类面是分隔面如图中 H1 和 H2 的位置，两个分隔面之间的都能够将样本正确的划分开[38]。其中，分类效果最好的分类面是图中 H 所在位置的分类面，即在分类间隔最中间的位

万方数据

置。线性判别函数的一般形式为：

$$g(x)=w^{T}x+b=0 \tag{3-7}$$

最优分类面的方程表示为：$w^{T}x+b=0$。满足函数的点为

$$y_i[w\cdot x_i+b]-1\geq 0, i=1,2,\cdots,n \tag{3-8}$$

满足上式，且使 2/||w||值最大的分类面为最优超平面。最优分类面问题还可以表示成如下约束优化问题，在式（3-8）的约束下求解函数，

$$\varphi(w)=\frac{1}{2}\|w\|^{2}=\frac{1}{2}(w^{T}w) \tag{3-9}$$

上式中最小值问题还可以转换为对偶问题，一般引入 Lagrange 函数进行求解，如下式：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^{2}-\sum_{i=1}^{i}\alpha_i(y_i(w_i+b)-1) \tag{3-10}$$

上式中，$\alpha_i=(\alpha_i,\cdots,\alpha_n)^{T}$ 为 Lagrange 乘子，$\alpha_i\geq 0$ 是 Lagrange 系数，为得到函数的最小值，对 w、b、α_i 进行偏微处理，使其等于 0。以上问题可以转化为凸二次规划的对偶问题[37]：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(x_ix_j)-\sum_{j=1}^{l}\alpha_j\\&s.t\sum_{i=1}^{l}y_i\alpha_i=0,\alpha_i\geq 0,i=1,2,\cdots l\end{aligned} \tag{3-11}$$

上述问题得到的解 $\alpha^{*}=(\alpha_1^{*},\cdots,\alpha_i^{*})^{T}$，每个$\alpha_i^{*}$的值都有一些样本点与之相对应。其中$\alpha^{*}>0$得到的样本点，即为支持向量。就可以得到最初问题的最优解，最优分类面的决策函数如下：

$$\begin{cases}f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^{*}y_i(x\cdot x_i)+b^{*})\\b^{*}=y_j-\sum_{i=1}^{l}\alpha_i^{*}y_i(x\cdot x_i)\end{cases} \tag{3-12}$$

根据上式可以得到，决策函数由支持向量决定，与α_i^{*}等于 0 的样本点没有关系。

在实际应用中，样本的类别复杂，能够将样本完全准确无误的区分开的分类平面是不存在的，因此线性可分满足不了现实问题，于是出现了线性不可分的解决方法[39]。

万方数据

3.3.2 线性不可分

线性不可分时，一个超平面没有将两类样本分开的性质，可以加入松弛变量 $\xi_i(\xi_i \geq 0, i=1,2,\cdots l)$，使最优分类面满足

$$y_i(w^T x_i + b) \geq 1 - \xi_i \tag{3-13}$$

此时得到的最优超平面为广义最优分类超平面，当 $0 < \xi_i < 1$ 时，样本点 x_i 可以被正确分类，而当 $\xi_i \geq 1$ 时，样本点 x_i 则未正确分类。因此，需要以下目标函数：

$$\psi(w,\xi) = \frac{1}{2} w^T w + C\sum_{i=1}^{n} \xi_i \tag{3-14}$$

式中，C 为正常值，称作惩罚因子。SVM 需要在原问题中分配一个代价函数，引入惩罚项 C，此时转化为以下目标函数：

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{l} \xi_i \\ &s.t. y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ &\xi_i \geq 0, i = 1,2,\cdots,l \end{aligned} \tag{3-15}$$

其中，此种方法适合样本近似线性可分的情况，对于复杂的样本分类问题，需要用复杂的超曲面代替分类超平面[40]。

其通过以下的对偶问题解决：

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j \\ &s.t. \sum_{i=1}^{l} y_i \alpha_i = 0 \\ &0 \leq \alpha_i \leq C, i = 1,\cdots,l \end{aligned} \tag{3-16}$$

其中 SVM 是通过一个映射函数φ，将训练样本变换到高维空间，使其在特征空间中线性可分，然后构造最优分类超平面，非线性变换通过核函数实现，核函数表示如下，

$$k(x_i \cdot x_j) = \phi(x_i) \cdot \phi(x_j) \tag{3-17}$$

其中，“ •” 表示内积。若 k（x_i • x_j）为半正定核，则问题为凸二次规划。线性不可分支持向量机与线性可分支持向量机的区别是α_i 的值受到了限制，若α_i*的值大于零，则 x_i 称为支持向量；若α_i*的值等于 C，则 x_i 称为边界支持向量；若α_i*的值在零与 C 之间，则 x_i 称为界内支持向量。由上式得到特解α*、b*，通过α*和核函数得到，最后得到决策函数：

万方数据

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^{*}y_i k(x_j\cdot x_i)+b^{*}) \tag{3-18}$$

最优分类超平面由α*>0 的样本决定，与其他无关。

3.3.3 核函数

核函数在支持向量机中的应用是一个显著特性，计算时核函数替代内积，能够应用到非线性问题中。核函数能够把不可分的低维数据映射到高维空间中进行线性分析，有效的解决了“维数灾难”问题。

在低维空间里，复杂的数据样本一般难以划分开，采用的方法是将数据向量集映射到高维空间，然后对数据集进行分类，此方法增加了计算的复杂度，增加了算法的难度。核函数的计算原理是，在原始空间先对向量集进行内积或者距离计算，然后将计算结果转换为非线性形式，最后将结果输入高维空间进行运算。核函数的加入使得大量复杂的运算在低维空间中进行，维度增加但是计算复杂度没有增多，巧妙的解决了“维数灾难”问题[19][41]。

在最优分类面中，满足 Mercer 条件的函数就能够实现非线性分类向线性分类的转换，问题变得容易解决，但计算难度并没有加大。

定理 1（Mercer）：对称函数 K（u,v）∈L2 能够以正系数 $a_k>0$ 展开成

$$k(u,v)=\sum_{k=1}^{\infty}a_k\psi_k(u)\psi_k(v) \tag{3-19}$$

上式中的充分必要条件是：

$$\int g^2(u)du<\infty \tag{3-20}$$

并且 g≠0，

$$\iint k(u,v)g(u)g(v)dudv\geq 0 \tag{3-21}$$

条件成立。

定义 1 核是一个函数 K，对于所有的 x,z∈X，满足：

$$K(x,z)=\langle\phi(x)\cdot\phi(z)\rangle \tag{3-22}$$

上式中，ϕ 运算是 x 到特征空间的映射，$F=\{\phi(x)\cdot\phi(z)\}$。

由于 SVM 是基于核函数之上的机器学习算法，核函数使得 SVM 的运算与样本的维度没有关系，简化了 SVM 的计算。转化的示意图如图 3-3 所示。

万方数据

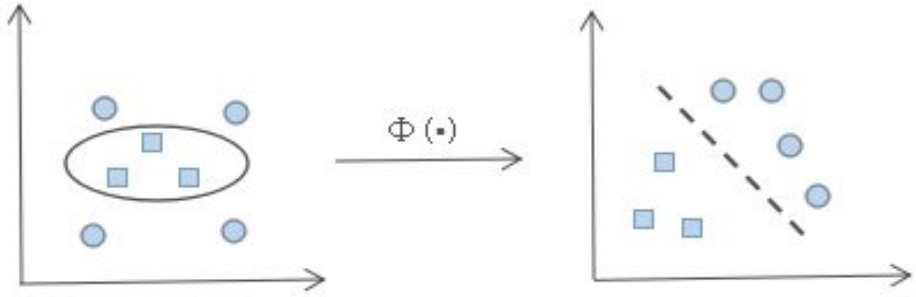

图 3-3 核函数非线性分类转化为线性分类

SVM 中常用的核函数包括以下核函数、线性核函数、Sigmoid 核函数和 RBF。

(1) 多项式核函数

函数的一般形式：

$$K(x,y)=(x\cdot y+1)^d \tag{3-23}$$

与之对应的 SVM 分类器的决策函数为（3-18），利用决策函数得到相关的系数，然后对待分类的向量进行运算，最后利用多项式得到最终值。

多项式核函数的计算与 SVM 的支持向量没有关系，而是与空间的维数和多项式的阶数相关。

多项式核函数在 x=1 时的图形如图 3-4 所示。

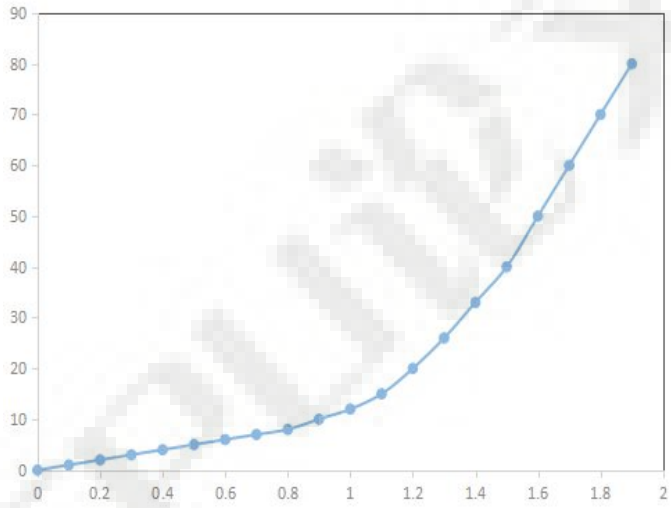

图 3-4 多项式核函数图像

(2) 径向基函数

径向基函数（Radial basis function，RBF），是沿径向对称的一种标量函数，高斯核函数是最常用的函数，其形式如下：

$$k(x,x_i)=\exp\left\{-\frac{\|x-x_i\|^2}{2\sigma^2}\right\} \tag{3-24}$$

万方数据

上式中，x_i为函数的中心，表示某点与核心点之间的距离，σ为宽度参数，表示对径向范围的控制。

当 x=0 时，径向基核函数分类器的图像，如图 3-5 所示。

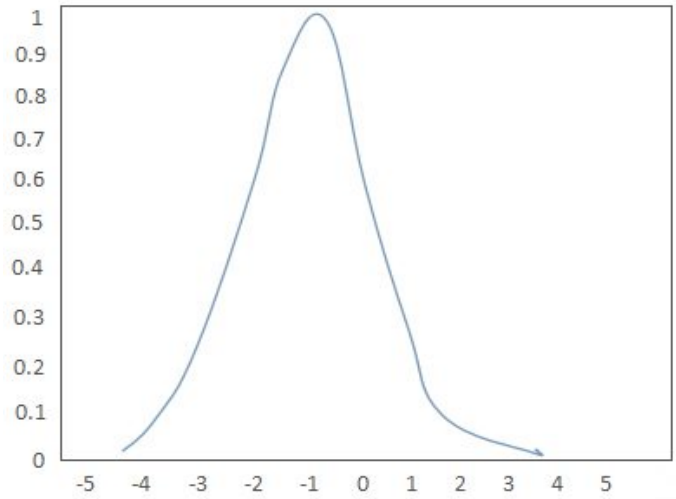

图 3-5 径向基核函数分类器

(3) Sigmoid 核函数

Sigmoid 核函数是一个感知器神经网络，网络的权值和节点数都是算法在计算中生成的，其公式如下：

$$K(x,y)=\tanh(v(x\cdot y)+c) \tag{3-25}$$

(4) 线性核函数

$$k(x,y)=(x,x_1) \tag{3-26}$$

以上四种核函数是比较常用的，多项式核函数在运算时，与空间维数和多项式阶数相关，当其的数量很大时，运算量也会相应的增加；由于高斯核函数有着较强的学习能力，在低维高维等情况下性能都很好，因此常用高斯函数作为核函数。

影响核函数的参数有惩罚因子 C 和核参数σ。C 控制对错分样本的惩罚程度，越大对错误的惩罚越重，泛化能力会降低。σ为函数的宽度参数，表示对径向范围的控制。当σ比训练样本间最小距离小很多时，则所有样本都是支持向量，都能正确分类；当σ比训练样本间最大距离大很多时，所有样本都分为一类，失去学习能力。因此，惩罚因子 C 和核参数σ的选取至关重要[22]。

3.4 实验结果与分析

在搜狗实验室下载的大量样本数据集中，包括经济、政治、环境、互联网和文化 5 大类，其中选取 350 篇作为实验样本，其中 250 篇作为训练数据集，100 篇作为测试集，分别对三种分类器进行测试。

万方数据

在实验中，文本表示采用特征选择时才用信息增益方法，对特征的信息增益值进行排序，选择前 2000 个为特征词；通过 TF-IDF 权值方法计算权值；最后对朴素贝叶斯、k 近邻和支持向量机分类器进行训练和测试。采用词频法和文档型概率进行计算，测试结果如表 3.1 和表 3.2 所示。

表 3.1 词频型测试结果

	朴素贝叶斯（%）			K 近邻（%）			支持向量机（%）		
	查准率	查全率	F_1 值	查准	查全率	F_1 值	查准率	查全率	F_1 值
经济	93.53	75	80.64	93.49	75	81	94.84	91	92.14
政治	91.23	100	96.42	81.25	98	86.73	91.24	91	93.65
环境	85.36	93	90.52	89.21	86	87.61	93.21	90	91.35
互联网	96.74	100	98.23	94.32	83	87.55	96.24	95	95.33
文化	100	96	98.25	85	100	94.53	96.34	98	97.57
平均值	93.372	92.8	92.812	88.654	88.4	87.484	94.374	93	94.008

表 3.2 文档型测试结果

	朴素贝叶斯（%）			K 近邻（%）			支持向量机（%）		
	查准率	查全率	F_1 值	查准率	查全率	F_1 值	查准率	查全率	F_1 值
经济	92.11	66	73.82	94.04	84	87.43	90.32	86	88.21
政治	86.75	96	91.12	91.36	100	87.61	87.32	98	93.62
环境	85.03	78	80.94	93.49	79	85.24	88.63	82	84.67
互联网	70.59	100	82.63	94.12	81	86.85	95.23	94	94.63
文化	100	96	98.25	75.09	100	84.47	100	98	99.27
平均值	86.896	87.2	85.352	89.62	88.8	86.32	92.3	91.6	92.08

根据表 3.1 能够得出，SVM 中经济和环境的 F1 值均高于朴素贝叶斯和 k 近邻，朴素贝叶斯算法中的政治、互联网和文化的 F1 值高于 SVM 的 F1 值，这是由于同一数据集可能属于不同的类别，但是平均值中 SVM 的 F1 值均高于另外两类，朴素贝叶斯算法的 F1 值高于 K 近邻算法。

万方数据

从表 3.2 能够得出，5 种类别中 SVM 的 F1 值均高于另外两种算法，经济、环境和互联网 K 近邻算法的 F1 值高于朴素贝叶斯算法。以文档频率为标准的结果中，从平均值分析，SVM 的 F1 值高于 K 近邻，K 近邻的 F1 值大于朴素贝叶斯。因为朴素贝叶斯算法的分类依赖于特征词的频率，所以在统计文档型概率方法时不如 K 近邻效果好，采用词频统计时效果优于 k 近邻算法。

通过实验，说明支持向量机是应用在文本分类中比较好的算法。常用的分类方法中，朴素贝叶斯算法和 k 近邻算法结构简单，易于实现。朴素贝叶斯算法要求属性是相互独立的，在现实中符合条件的很少；K 近邻算法，类别的确定与相邻的类别有关，k 值不易确定而且随机取值会增加算法时间，影响算法的效率。

支持向量机的优点：在解决小样本、非线性问题和高维度空间问题时展现出了很大的优越性；避免了过学习和局部最优问题；低维数据向高维空间转换时，计算的复杂度没有增加，有效的克服了“维数灾难”；提高了泛化能力，是统计学中较为实用的算法。

但是 SVM 也存在一些缺点，在实际应用中，选择合适的函数比较困难；SVM 的特点是参数对其分类性能的影响很大，但是选取合适的参数又比较困难，很多都是依靠经验来选择，缺少理论基础；SVM 适用于二分类问题，在多分类问题上效果不太好[23]。对 SVM 中的参数进行优化是本文的重点研究内容之一，本文将在第 4 章进行详细介绍。

3.5 本章小结

本章主要介绍了文本分类中常用的算法，包括朴素贝叶斯算法、k 近邻算法和支持向量机，并对三种算法通过实验进行了比较和分析。实验利用文档统计和词频统计概率方法，通过查准率、查全率和 F 测量值分析了各个算法的分类性能。结果表明，两种概率模型下，SVM 都明显高于其他两种算法。最后总结了支持向量机的缺点和目前存在的问题。

万方数据

第 4 章 改进的 SVM 参数优化方法

4.1 SVM 参数

SVM 在训练样本数量有限的情况下，分类识别准确率和分类推广能力表现出了很好的效果，在处理非线性和高维空间问题时也具有一定的优势。但是 SVM 对核函数、参数的选择比较敏感，算法的性能受设置的参数影响很大。

SVM 中对其算法性能影响较大的因子包括核函数、核函数的参数以及误差惩罚因子。不同的核函数对于算法的分类效果影响不明显，其中惩罚因子 C 和核参数 σ 的选值对算法性能影响较大。

（1）惩罚因子

惩罚因子 C 的值越大，代表对经验误差的惩罚越大，机器学习的复杂度越大，经验风险值就越小。若 C 的值为无穷大，则代表约束条件全部都符合，表示训练样本必须准确无误的分类。每个特征子空间中至少有一个 C 值使得 SVM 的推广能力最强，C 值过大或者过小 SVM 的经验风险和推广能力都不会改变。

（2）核参数

SVM 一般的核函数有多项式核函数、线性核函数、Sigmoid 核函数和高斯核函数（简称 RBF）[24]。由于高斯核函数具有较强的学习能力，函数识别能力强，在低维高维等情况下均适用，因此常用高斯函数作为 SVM 的核函数。

高斯函数中σ为函数的宽度参数，当σ比训练样本间最小距离小很多时，σ的值趋于 0，则表示所有的样本都是支持向量，均能正确分类，造成“过度拟合”的现象，导致推广能力降低；当σ比训练样本间最大距离大很多时，样本分类的错误率很低，但是所有样本均分为一类，使得分类器失去学习能力，分类效果降低[25]。

4. 2 SVM 参数优化方法

SVM 的参数选取问题，是一个参数优化的过程。传统的 SVM 参数选取方法包括交叉验证法、网格搜索法，这两种算法虽然原理简单，但是计算量偏大。下面对主要的算法进行介绍。

4.2.1 交叉验证法

交叉验证法（Cross Validation，简称 CV）是基于统计学的方法，基本思想是将文本数据分为两组，一组成为训练样本，另一组成为测试样本。首先利用训练样本对分类器进行学习训练，经过学习训练优化得到最合适的分类器参数，然后通过测

万方数据

试样本对样本模型进行测试，最后将测试样本的分类结果作为评价分类模型分类效果的标准[26]。

交叉验证法的一般方法包括 Holdout 验证法、K 重交叉验证法和留一验证法。主要内容如下：

（1）K 重交叉验证

K 重交叉验证（K-fold Cross Validation，K-CV），基本原理是将训练文本平均分成 k 份，把其中的 k-1 份分配为训练数据子集，将最后一份作为测试数据子集，将此方法重复进行 k 次，即每一份都担任过训练数据子集，然后根据 k 次的平均值估算误差，选取结果最优的一组作为参数。

（2）Holdout 验证法

Holdout 验证法的原理是从原始数据集中随机的选取若干样本作为测试样本，其余的成为训练样本，选取测试样本的数量比原始数据集的总量要少，一般少于三分之一。

4.2.2 网格搜索法

网格搜索法优化的是惩罚因子 C 和核函数参数，假设惩罚因子的范围 $C\in[c_1,c_2]$，变化步长设定为 c_s，核函数参数σ范围$\sigma\in[\sigma_1,\sigma_2]$，变化步长是$\sigma_s$。对每对参数进行优化训练，选取最佳的一对作为参数模型[28]。

网格搜索法是针对每对参数进行优化，能够将所有参数遍历到，得到的最佳参数可靠性较高；此方法涉及到惩罚因子和核函数两个参数，在进行训练能够平行计算，进而可以节省时间。但是网格搜索法也有一些缺陷，因为要遍历每对参数，所以计算量会很大。

以上三种算法都是比较传统的方法，有其优越性，也存在一些缺点，因而提出了一些改进的算法，提高 SVM 的算法性能。

4.3 萤火虫算法优化 SVM 参数

4.3.1 标准萤火虫算法

萤火虫算法（Firefly Algorithm，简称 FA）是由英国剑桥教授 Yang 于 2008 提出的一种新兴的群智能优化算法。萤火虫个体利用自身发出的光与其他个体进行联系，光的强度会随着距离变大会逐渐减小，个体间的距离越大传递信息的准确性会越低[26][45]。

萤火虫算法的主要原理是把每个荧火虫个体看成空间的点，发光强的萤火虫个体会将发光弱的萤火虫吸引过来，发光弱的个体向发光强的个体移动，吸引度不同

万方数据

移动的距离也不同，移动后的萤火虫自身的发光强度和吸引度得到更新，最后萤火虫聚集在亮度最强个体的位置，完成目标的优化[27]。

萤火虫算法的实现基于以下三个要求：

（1）假设萤火虫个体不存在性别的差异，都能够相互吸引；萤火虫的亮度与其所处的位置有关，位置越好亮度越大。

（2）萤火虫个体间的相互吸引的程度，与萤火虫间的相对亮度以及距离紧密联系。亮度强的个体吸引同方向的个体越多，对距离越远的个体吸引度越弱，亮度相同的个体会做随机运动。

（3）在实际应用中，萤火虫个体的发光强度由目标函数决定，在特定范围内与目标函数成比例。

4.3.2 算法原理

萤火虫算法是一种智能搜索算法，算法原理是模拟萤火虫个体间的相互吸引行为和萤火虫位置不断更新的过程。因此，算法的求解过程就是寻找发光最亮的萤火虫个体。

算法中两个重要的参数为萤火虫的自身亮度和个体间的吸引度，发光强的个体会吸引发光弱的个体，亮度越大吸引力越强，亮度弱的个体就会向亮度强的个体移动，然后聚集在最亮个体的位置，亮度相同的个体会做随机运动，最后亮度最强的萤火虫即为最优解[28][46]。

（1）萤火虫的荧光亮度与目标函数值相关，萤火虫的位置越好，亮度越强，代表的函数值越优。个体间的相互吸引会随着距离的变化而改变，距离越大吸引度越弱，在传输过程中萤火也会被传播介质所吸收，吸引度的大小跟介质吸引因子相关联[29]。萤火虫的荧光亮度用下式表示：

$$I = I_0 * e^{-\gamma r_{ij}} \tag{4-1}$$

式中 I0 代表萤火虫自身亮度，距离为 0 时，萤火虫的亮度最强，即 I0；γ为光强吸收系数，理论上取[0, ∞)，一般问题中取值为[0.01,100]；r_{ij} 为萤火虫个体 i 与萤火虫个体 j 的欧式距离，计算公式如下：

$$r_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})} \tag{4-2}$$

式中 x_i 和 x_j 表示萤火虫个体 i 和 j，d 表示总的空间向量维数，$x_{i,k}$ 表示萤火虫 i 在空间向量中的第 k 个分量。

（2）萤火虫的吸引度，表示个体之间的相互吸引度。

$$\beta = \beta_0 * e^{\gamma r_{ij}^m} \tag{4-3}$$

万方数据

式中 β_0 为最大荧光亮度位置的吸引度，即距离为 0 时的吸引度；γ为光强吸收系数，萤火虫的荧光会随着距离与传播介质吸收系数的增大逐步减小，对其他个体的吸引也相对减少[30]。

（3）萤火虫位置更新，萤火虫 i 被萤火虫 j 吸引后，位置更新公式：

$$x_i(t+1) = x_i + \beta \times (x_j - x_i) + \alpha \times (rand - 1/2) \qquad (4\text{-}4)$$

式中，x_i 与 x_j 表示萤火虫个体 i 和个体 j 的位置；α表示步长因子，为常数因子，一般取值为[0,1]；rand 是均匀分布的随机因子，一般取值为[0,1]。

萤火虫算法的具体实现原理，假设存在三个萤火虫个体 A、B、C，每个的搜索半径为 r_A，r_B，r_C。萤火虫 B 在萤火虫 A 的搜索半径内，A 的搜索半径要比 B 的半径大，当 A 的亮度强于 B 的亮度时，个体 B 向个体 A 移动，当 A 的亮度弱于 B 的亮度时，则 A 向 B 移动。因为萤火虫 C 均不在 A 和 B 的搜索范围内，所以三者的亮度强弱不影响个体 C 的移动方向，具体位置如图 4-1 所示。

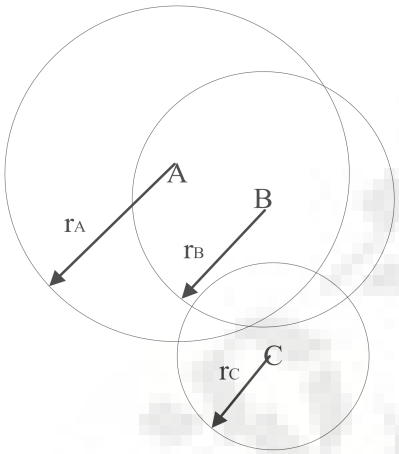

图 4-1 萤火虫算法搜索图

萤火虫算法存在一些优点和缺点，由于萤火虫算法原理是模拟萤火虫的发光特性，所以算法结构简单，需要调节的参数少；发光亮的萤火虫个体会吸引发光弱的个体，萤火虫会向亮度强的个体移动，通过位置的不断更新迭代可以得到全局或局部的最亮个体，即目标的最优解[32]。

萤火虫算法结构简单，涉及到的参数个数少，易于计算。萤火虫算法优化 SVM 的具体实现过程的流程图，如图 4-2 所示：

万方数据

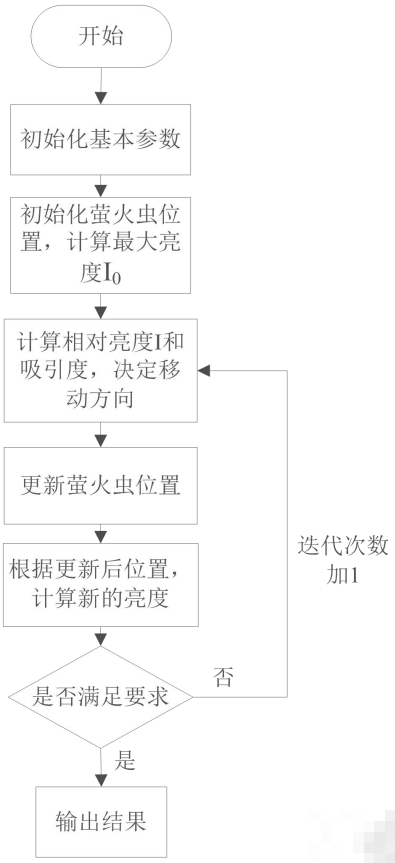

图 4-2 萤火虫算法流程图

如果萤火虫的亮度一样，萤火虫的移动方向会不定变化，个体会做随机运动很难找到最优解；若萤火虫个体的搜索半径内没有更亮的个体，萤火虫做随机运动，导致算法的收敛速度和搜索能力下降。

4.3.3 萤火虫算法优化 SVM 参数

标准萤火虫算法优化 SVM 参数过程如下：

（1）初始化各参数，萤火虫的数目 m，最大荧光亮度的吸引度β_0，光强吸收系数γ，步长因子α_0，可搜索的范围$[x_{min}, x_{max}]$；萤火虫的初始位置，最大迭代次数 n。X 轴的取值范围为$[C_{min}, Cmax]$，y 轴的取值范围为$[\sigma_{min},\sigma_{max}]$。

（2）根据 libsvm 包，计算惩罚因子 C 和核参数σ的初始值。

（3）由更新后的萤火虫位置，重新计算萤火虫的荧光亮度和吸引度，确定萤火虫的移动方向。

（4）根据公式（4-4）更新萤火虫位置，对最优萤火虫的位置进行扰动。

（5）计算更新后萤火虫的亮度。

（6）达到最大搜索次数或者满足搜索精度时，则转为下一步；否则搜索次数增加，回到步骤（3）重新搜索。

万方数据

（7）输出全局极值点和最优解。

4.4 改进的萤火虫算法优化 SVM 参数

4.4.1 改进的萤火虫算法

标准的萤火虫算法本身具有很大的优势，结构简单易于操作，调节参数少。但是萤火虫算法本身也存在着一些缺陷，算法容易陷入局部的最优，收敛过早。在标准萤火虫算法中，萤火虫位置更新时，步长因子α的值是固定的。若α取值过大会导致算法后期容易陷入局部最优，α取值过小，算法早期的搜索能力将较弱[33][34]。

为了平衡算法进化过程中的搜索能力与收敛速度，改变原来固定的α值，对α进行了变形处理。在步长因子中加入了迭代次数变量，调整α为线性降低的数值，随着迭代次数的增加步长逐渐减小，避免了步长不稳定的现象，本文改进后的萤火虫算法记为 IFA。调整后的α如下式：

$$\alpha = \alpha_0 * \frac{\left\| x_i - x_{best} \right\|}{d_{\max}} * \frac{n}{i} \tag{4-5}$$

式中α_0为初始化步长因子，x_{best}为最优萤火虫的位置，$\|x_i-x_{best}\|$表示当前萤火虫与最优萤火虫的空间距离，可用欧式距离代表；d_{max}表示最优萤火虫个体与邻域内萤火虫的最大距离[35]。

更新公式后，在算法初期，α的值比原来增大，步长的搜索范围也变大，搜索能力增强；随着迭代次数的增加α线性减小，在算法后期搜索步长逐渐减小，收敛速度增加，避免了后期算法收敛速度过慢的问题。

为了避免搜索步长过大或者过小，保证更新后位置的萤火虫在设定的搜索范围内，对步长范围进行了限制，对于超出搜索范围的萤火虫位置按照下式处理：

$$x_i \begin{cases} x_{\min}, x_i < x_{\min} \\ x_{\max}, x_i > x_{\max} \end{cases} \tag{4-6}$$

当搜索长小于等于最小值时，则把搜索范围的最小值作为步长，当搜索步长大于等于最大值时，则把搜索范围的最大值作为步长，保证步长不会过大或者过小，高效率的进行搜索。

本文提出的萤火虫算法，根据萤火虫的搜索原理对搜索步长因子进行了改进。标准的萤火虫算法中步长因子是一个常数，在萤火虫的位置更新过程中α起到了很大的作用。改进后的算法将其调整为随迭代次数增加线性减少的变量，算法改进后，在早期搜索步长变大，进而搜索能力增强；算法后期，步长逐渐减小，使收敛速度加快，减少了算法的搜索时间。

万方数据

4.4.2 SVM 参数优化

根据萤火虫算法的原理，基于改进的萤火虫算法对 RBF 参数优化过程如下：

（1）初始化各参数，萤火虫的数目 m，最大荧光亮度的吸引度β_0，光强吸收系数γ，步长因子α_0，可搜索的范围$[x_{min}, x_{max}]$；萤火虫的初始位置，最大迭代次数 n。X 轴的取值范围为$[C_{min}, C_{max}]$，y 轴的取值范围为为$[\sigma_{min},\sigma_{max}]$[36][45]。

（2）根据 libsvm 包，计算惩罚因子 C 和核参数σ的初始值。

（3）由更新后的萤火虫位置，重新计算萤火虫的荧光亮度和吸引度，确定萤火虫的移动方向。

（4）根据改进后的 FA 更新萤火虫位置，对最优萤火虫的位置进行扰动。

（5）计算更新后萤火虫的亮度。

（6）达到最大搜索次数或者满足搜索精度时，则转为下一步；否则搜索次数增加，回到步骤（3）重新搜索。

（7）输出全局极值点和最优解，程序结束。

根据改进的萤火虫算法优化核函数参数的过程，得出了核函数参数优化过程流程图，如图 4-3 所示。

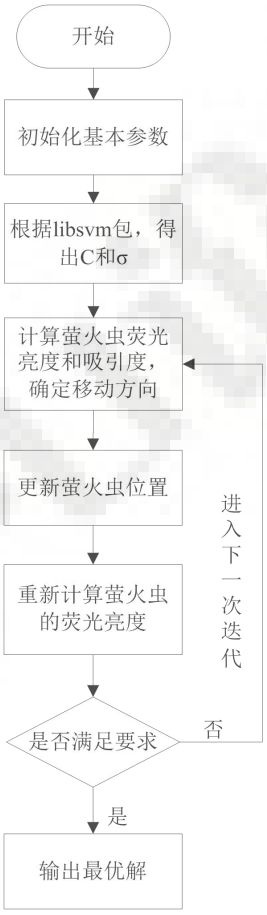

图 4-3 萤火虫算法优化 SVM 参数流程图

万方数据

4.4.3 实验对比与分析

为了验证改进萤火虫算法的性能效果，选取了 Ackley 函数、Sphere 函数和 Rosenbrock 函数进行仿真测试。种群数量设置为 25，光吸收系数 γ 为 1，最大荧光亮度的吸引度 β_0 为 2，步长因子 α_0 为 0.2，维数设为 10，最大迭代次数根据函数的不同设置为[100,1000]。表 4-1 是选取的函数表达式。

表 4-1 测试算法性能的经典函数

函数名	函数表达式	取值范围	最优解
Ackley	$f(x)=-20e^{(-0.2\sqrt{\frac{1}{10}\sum_{i=1}^{n}x_i^2})}-e^{\frac{1}{10}\sum_{i=1}^{n}\cos 2\pi x_i}+20+e$	[-100,+100]	0
Sphere	$f(x)=\sum_{i=1}^{n}x_i^2$	[-100,+100]	0
Rosenbrock	$f(x)=\sum_{i=1}^{n-1}(100(x_{i+1}-x_i^2)^2+(x_i-1)^2)$	[-30,30]	0

(1) Ackley 函数的仿真实验

在 Ackley 函数的仿真实验中，为了验证本文改进算法的效果，迭代次数均设置为 500 次，每个算法实验进行 10 次，最后得到的数据取平均值，实验数据如表 4-2 所示。

表 4-2 Ackley 函数数据表

算法名称	总次数	总时间（s）	最优解次数	最优解
FA	500	16.011340	500	8.8818e-15
IFA	500	4.078083	62	2.2204e-15

在 MATLAB 中两个算法对 Ackley 函数进行实验，得到两个搜索过程的曲线图，每次曲线图会有些不同，选取的是较为稳定时的曲线，如图 4-4 所示。

万方数据

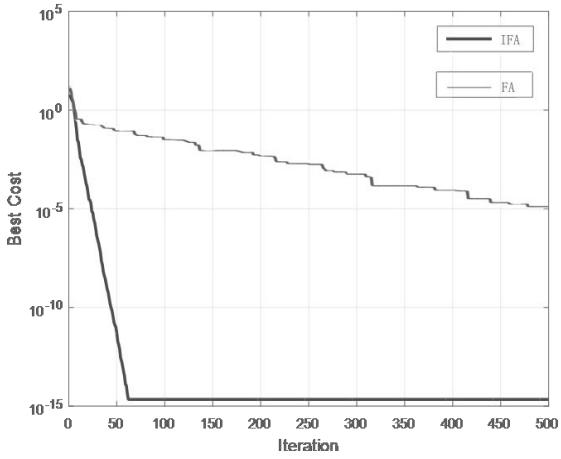

图 4-4 Ackley 函数的曲线对比图

（2）Sphere 函数仿真实验

在 Sphere 函数的仿真实验中，为了验证本文改进算法的性能，迭代次数均设置为 1000 次，每个算法实验进行 10 次，最后得到的数据取平均值，实验数据如表 4-3 所示。

表 4-3 Sphere函数数据表

算法名称	总次数	总时间（s）	最优解次数	最优解
FA	1000	18.561770	1000	6.6725e-21
IFA	1000	13.571364	644	0

在 MATLAB 中两个算法对 Sphere 函数进行实验，得到两个搜索过程的曲线图，选取的是较为稳定时的曲线，如图 4-5 所示。

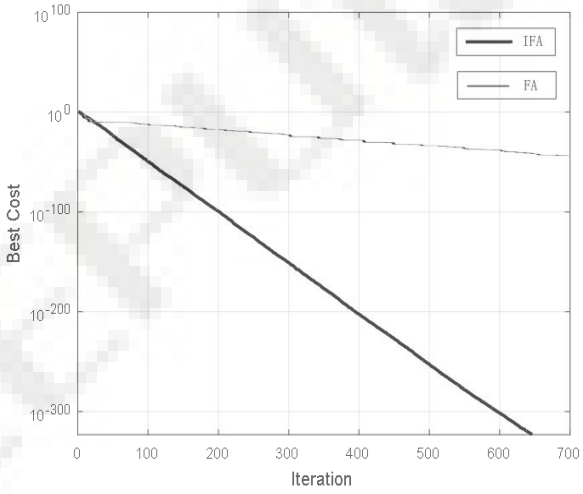

图 4-5 Sphere 函数的曲线对比图

（3）Rosenbrock 函数仿真实验

万方数据

在 Rosenbrock 函数的仿真实验中，为了验证本文改进算法的效果，迭代次数均设置为 500 次，每个算法实验进行 10 次，最后得到的数据取平均值，实验数据如表 4-4 所示。

表 4-4 Rosenbrock函数数据表

算法名称	总次数	总时间（s）	最优解次数	最优解
FA	500	15.654240	500	8.9534e-07
IFA	500	10.217560	500	4.7037e-08

在 MATLAB 中两个算法对 Sphere 函数进行实验，得到两个搜索过程的曲线图，选取的是较为稳定时的曲线，如图 4-6 所示。

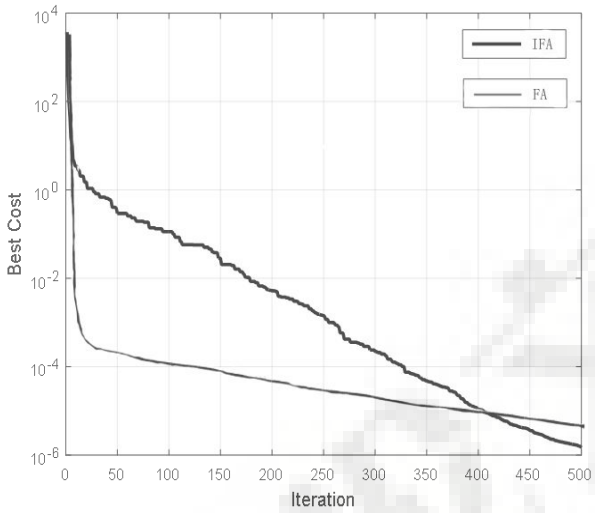

图 4-6 Rosenbrock 函数的曲线对比图

由仿真实验的表格与对比曲线图可知，在 Ackley 函数中，IFA 在迭代过程中快速收敛，在次数为 62 得到最优解；FA 的曲线平缓，在迭代次数为 500 次时，没有得到最优解。IFA 的搜索时间为 FA 的 1/4。在 Sphere 函数中，IFA 在迭代次数为 644 时得到最优解，FA 在迭代 1000 次时仍没有得到最优解。在 Rosenbrock 函数中，虽然两种算法最优解相差很少，但是 IFA 在早期快速收敛，趋于最优解。实验结果表明，IFA 前期搜索步长增加，使得全局搜索能力增强，后期随着迭代次数增加，步长线性减少，收敛速度加快，收敛趋于稳定，得到最优解，验证了 IFA 算法的优越性。

4.5 本章小结

本章介绍了传统的优化 SVM 核参数的方法，针对传统方法的不足，引入了萤火虫智能算法。萤火虫算法具有结构简单易于处理的优势，但也存在着前期搜索能

万方数据

力不强，后期易陷入局部最优的问题，本文对萤火虫的位置更新公式进行了改进，加入了迭代次数因子，算法在早期的搜索能力增强，后期的收敛速度加快，避免算法后期陷入局部最优。通过对标准萤火虫和改进的萤火虫算法进行仿真实验验证了改进萤火虫算法的高效性，并将改进后的萤火虫算法用于优化 SVM 参数中，得到一种改进的优化 SVM 参数的方法。

万方数据

第 5 章 文本分类实验及结果分析

5.1 实验说明

本文实验在 MATLAB R2015a 环境下进行，在 Windows 环境下实现，计算机的系统配置如下表 5-1 所示：

表 5-1 计算机配置

计算机结构	配置
操作系统	Windows 7
处理器	Intel(R) Core(TM) i5-4558u cpu@2.4GHz
内存	8.00GB
硬盘	500GB
系统类型	64 位操作系统

实验的软件环境是 Matlab 2015a，开发包选用的是 LIBSVM 是台湾林志仁教师开发的模式识别与回归软件包，本实验在 Windows 环境下进行，所以可以对开发包自带的编译程序进行修改，以达到实验的标准。

实验中 libsvm 包的数据格式如下：

<label><index>:<value1><index2>:<vaule2>…

<label>标签代表类别标识，一般是整数形式；<index>的数值从 1 开始，都是整数；<value>是进行训练的数据，是分类数据中的特征值。

5.2 文本分类测试语料

语料库是指通过语言研究和电子信息保存下来的材料，其内容可以是口语形式的也可以是书面形式的，经过后期的处理后，成为研究语言的材料[47]。

测试的形式有两种，封闭测试和开放测试。封闭测试指测试的数据集是训练集中的一部分或者全部；开放测试指的是测试数据集不包括在训练集中。两种方式各有优缺点，本文使用的是开放测试[48]。

本文选用的分类数据集同样来自搜狗实验室，从 2006 年至今，经过了人工编辑和整理，反复的补充和更新，数据集类别丰富分类专业。语料库一共有 10 个类别，每个包含 7999 篇文档，本文选择其中的经济、政治、环境、互联网和文化五个类别进行实验，其中每个类别选择 1000 篇，700 篇作为训练集，300 篇作为测试集。

万方数据

语料库中，类别文件夹的名称以不同编号命名，文本格式为 TXT 文件，文件名以数字编号命名，每个类别中的文档均从数字 10 开始编号，保证了分类的公平性。文档的形式如图 5-1，5-2，5-3 所示：

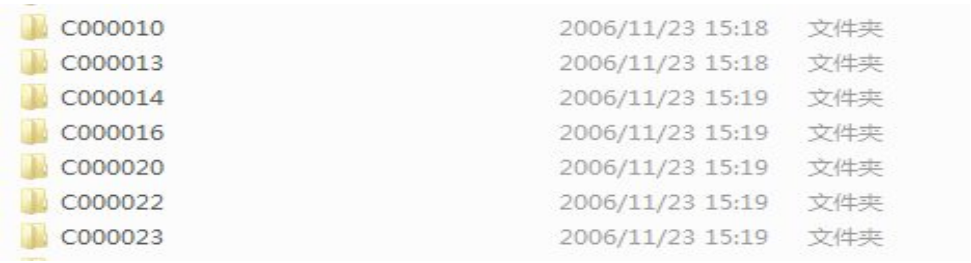

图 5-1 文本类别

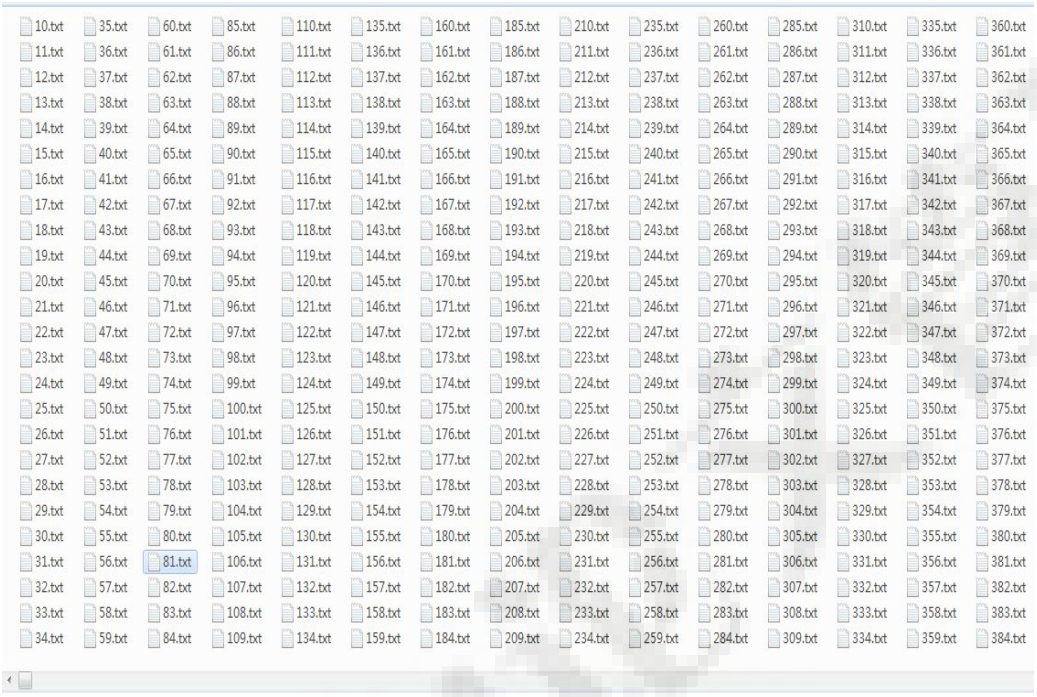

图 5-2 文档形式

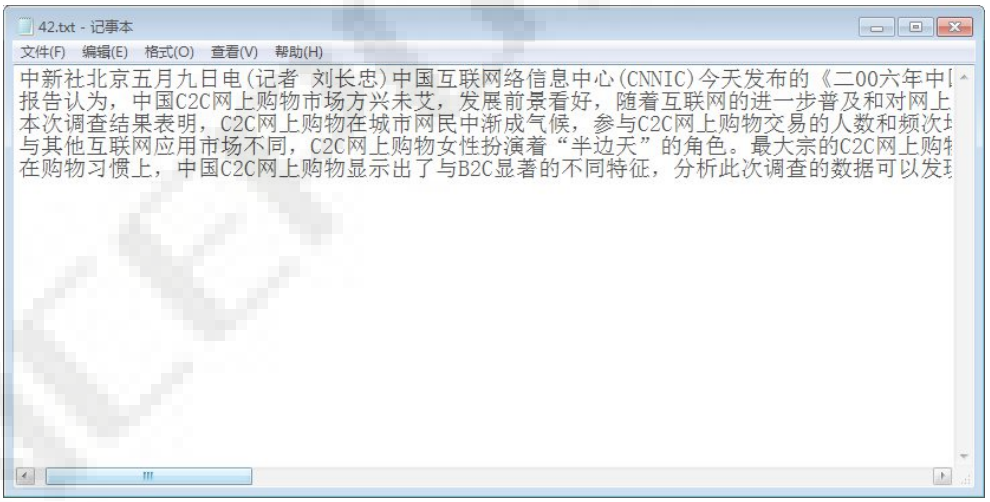

图 5-3 文本内容

万方数据

5.3 文本分类实验过程

实验的具体过程按照文本分类的顺序进行，主要步骤包括文本预处理、文本表示、特征处理、分类器和评价分类效果。

文本预处理就是提取信息关键字，过滤噪声，减少对的影响分类，提高分类准确率。本次实验文本预处理建立在 ICTCLAS 基础之上，ICTCLAS（Institute of Computing Technology, Chinese Lexical Analysis System）汉语词法分析系统，是中国科学院计算机技术研究所多年研究的科研成果，其中它包括的功能有很多，比如中文分词、词性标注等，还能够支持用户字典。本实验主要应用其中文分词的功能[49]。

文本表示，计算机无法直接处理文字，需要将数据转化为计算机能够处理的形式，本次实验采用向量空间模型，此模型结构简单、易于操作计算；计算特征词之间的相关程度，忽略了语法、语义等复杂关系，减少了计算的工作量，计算结果更客观可信。

特征处理，经过文本预处理和文本表示之后，进行特征提取，提取后文本向量的空间维数较大，需要对文本向量进行降维处理，本文使用信息增益方法进行降维处理。文本中每个特征词对文本分类的贡献度不同，因此要将特征项赋于不同的值，本实验采用 TF-IDF 权重法，统计词语在文档中的出现频率，频率越高贡献率越大。

文本分类器使用 SVM 算法，在选取 SVM 参数时，通过改进的萤火虫算法进行选取。 具体步骤如下：

(1) 将实验所需的数据集转化为 MATLAB 要求的格式并导入 MATLAB。

(2) 选择核函数的类型，本实验选择 RBF 作为支持向量机的核函数。

(3) 将改进后的萤火虫算法对核函数参数进行选择。

(4) 将得到的参数对训练集进行训练[50]。

(5) 将得到的支持向量机模型对文本进行分类。

5.4 实验结果与分析

将改进后的 SVM 分类器（I-SVM）与标准的 SVM 分类器（SVM）进行比较分析实验结果，验证本文提出方法的可行性和高效性。分类效果通过统计文档的查准率（P）、查全率（R）和 F 测量值（F1）进行比较。

改进后的萤火虫的参数设置，光吸收系数γ为 1，最大荧光亮度的吸引度β_0为 2，最大迭代次数为 300，步长因子α_0设置为 0.2。为了提高实验的准确性、真实性，在选择σ和 C 时，将每种类别平均进行 5 次，得到平均值再应用于文本分类中，实验结果如表 5-2。

万方数据

表 5-2 参数值

	I-SVM		SVM	
	C	σ	C	σ
经济	3703	0.15	3417	0.23
政治	4042	0.19	3820	0.31
环境	4404	0.08	4143	0.14
互联网	4321	0.02	4021	0.07
文化	4673	0.03	4361	0.09

表 5-3 结果对比图

	I-SVM			SVM		
	查准率	查全率	F_1 值	查准率	查全率	F_1 值
经济	94.37	91	92.78	90.31	83	88.32
政治	92.83	92	93.83	85.67	90	87.83
环境	95.32	92	94.15	87.41	85	86.37
互联网	97.83	98	98.42	94.31	86	89.63
文化	98.84	98	97.34	87	98	95.73
平均值	95.838	94.2	95.304	88.94	88.4	89.576

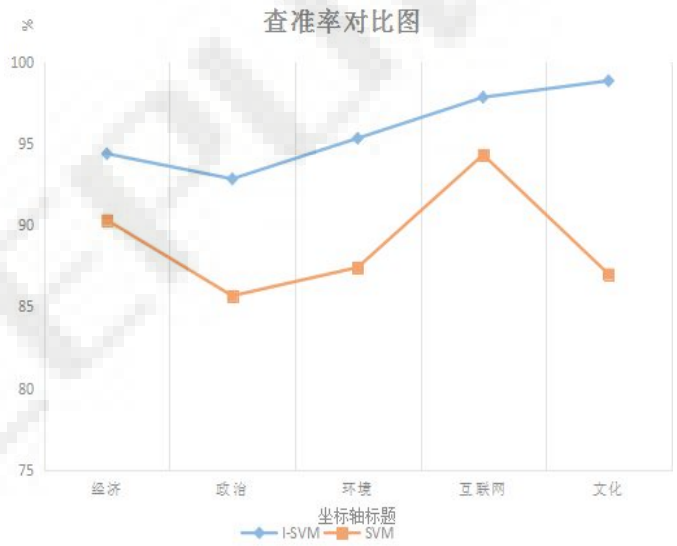

图 5-4 查准率对比图

万方数据

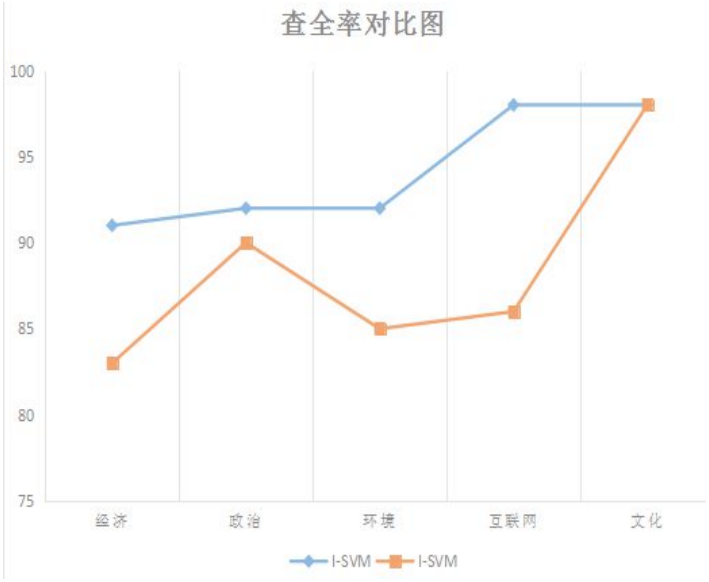

图 5-5 查全率对比图

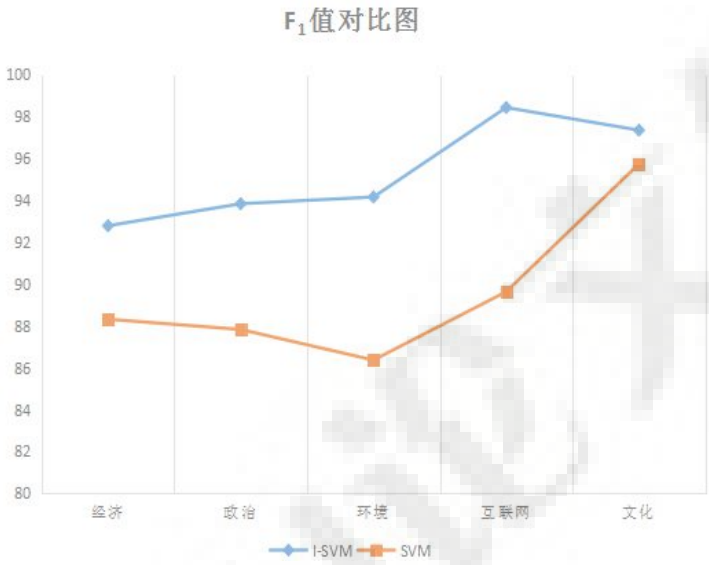

图 5-6 F_1 值对比图

根据表 5-2 可知，I-SVM 中惩罚因子 C 的值均比 SVM 中的值大，从经济到文化 5 个类别，C 的值相差 200-300，C 的主要功能是调整经验风险的比例，提高推广能力，I-SVM 中 C 的值相对较大，说明对经验惩罚越大，经验的风险就越小，既保证了 SVM 的推广能力，又保证了风险的比例；相反，在每个类别中，I-SVM 中 RBF 的σ值均比 SVM 中的值小，σ的主要作用是控制径向的范围，σ的值越小支持向量的个数越多，正确分类的个数越多，但是σ的值控制在一定范围内，所以保证了分类器的学习能力。

根据表 5-3 可知，I-SVM 算法的查准率、查全率和 F1 值均高于 SVM 的值，I-SVM 中文化的查准率比 SVM 中的查准率高 11.84%，互联网中的查全率前者比后者高 12%，F1 的值前者比后者高 8.79%，三个评价标准前者的值均高于后者，验证了改

万方数据

进后萤火虫算法优化 SVM 后得到的参数，有效的提高了分类器的准确率，说明本文方法是有效可行的。

两种算法实验过程中，每个分类类别需要的时间也不同，所得结果如图 5-7 所示。

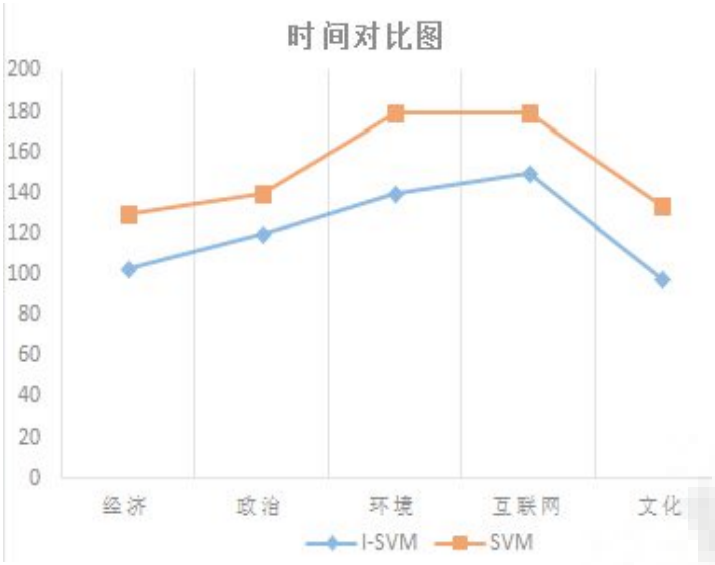

图 5-7 两种算法的时间对比图

根据图 5-7 可知，I-SVM 的分类时间均要少于 SVM 所需的时间，I-SVM 在环境分类中需要的时间要比 SVM 中需要的时间要少 50s。利用本文改进萤火虫算法对 SVM 进行参数选择后，得到的参数应用文本分类，分类的速度提高，分类的时间减少，有效的证明了本文提出方法能够加快文本分类速度，减少分类时间。

5.5 本章小结

本章主要介绍了文本分类的实验，改进后萤火虫算法对 SVM 的参数选取，将改进后的分类器应用于文本分类，并与标准 SVM 分类器进行对比。介绍了实验的软硬件环境、实验所需的测试数据和详细的实验过程，并对实验结果进行了分析。实验验证了本文改进的 SVM 算法有效的提高了文本分类的准确率和速度，加快了分类的时间，说明了算法的可行性。

万方数据

第 6 章 总结与展望

现在网络传递信息的能力已经超越了我们的想象力，信息海量增长，但是如何从海量信息中获取有用信息却困扰着人们。此时文本分类应运而生，文本分类可以高效的组织和管理信息，实现快速、准确的定位信息，有效的解决了信息杂乱无章的问题。

6.1 本文总结

本文主要介绍了文本分类的研究背景和研究意义，文本分类的一般过程和相关的理论与技术，本文研究的主要内容如下：

（1）对文本分类中常用的算法 SVM、朴素贝叶斯算法和 K 近邻算法进行了比较，通过研究和实验，实验结果表明 SVM 是文本分类中分类速度和准确率最高的算法。

（2）引入了萤火虫算法，萤火虫算法结构简单、易于操作，调节参数少。但是萤火虫算法存在着容易陷入局部最优和搜索能力较弱的缺点。文本对萤火虫算法进行了改进，通过实验验证了改进后的萤火虫算法性能增强，早期搜索能力增强，后期收敛速度加快。

（3）传统的 SVM 参数选取方法虽然原理简单，但是计算量偏大，本文将改进后的萤火虫算法应用于SVM参数优化中，并将优化后的参数应用于训练SVM模型。

（4）通过实验进行比较，实验结果验证了改进后的 SVM 算法应用于文本分类后，在分类精准度和速度方面都高于标准的 SVM，验证了提出的新算法的可行性和高效性。

6.2 研究展望

由于个人现在掌握的知识水平和时间的限制，论文工作还需要进一步深入的研究，主要包括：

（1）本文改进的算法是基于萤火虫算法，虽然萤火虫算法结构简单，但在实验过程中需要多个软件的支持，加重了计算机的负担，应该研究出更方便快捷的选取 SVM 参数的方法。

（2）由于实验使用的是 pc 电脑，实验处理的数据集数量有限，对大量的数据集无法进行实验，现实中的数据集是庞大的，所以需要更好的硬件设施来处理大量数据问题。

万方数据

（3）特征处理是文本分类的核心步骤，但是当前的算法结构复杂而且计算量大，应进一步研究更好的特征处理方法，提高文本分类的效率。

万方数据

参考文献

[1] 林振飞.基于混合特征的中文文本分类研究[D]. 东北大学, 2012
[2] 张振浩.中文文本自动分类关键技术研究及实现[D]. 浙江理工大学, 2013
[3] 胡泽铭.数字界面视觉信息认知 ERP 实验研究[D]. 东南大学, 2015
[4] 刘海峰,苏展,刘守生.一种基于词频信息的改进 CHI 文本特征选择[J]. 计算机工程与应用, 2013,(49)22, 110-114
[5] 李学学.基于数据预处理和回归分析技术的数据挖掘算法及其应用研究[D]. 兰州交通大学,2014
[6] 邸锦. 基于支持向量机的文本分类问题的研究[D]. 北京交通大学, 2008
[7] 徐晓明.SVM 参数寻优及其在分类中的应用[D]. 大连海事大学, 2014
[8] 李琼,陈利.一种改进的支持向量机文本分类方法[J]. 计算机技术与发展, 2015, 25(5): 78-82
[9] 王小青. 中文文本分类特征选择方法研究[D]. 西南大学, 2010
[10] 刘依璐. 基于机器学习的中文文本分类方法研究[D]. 西安电子科技大学, 2015
[11] 杨海. SVM 核参数优化研究与应用[D]. 浙江大学, 2014
[12] 杜芳华. 基于半监督学习的文本分类算法研究[D]. 北京工业大学, 2014
[13] 辛竹.文本分类中的特征提取算法研究与改进[D]. 北京邮电大学, 2014
[14] 郑俊飞.文本分类特征选择与分类算法的改进[D]. 西安电子科技大学, 2012
[15] 张岩.基于 SVM 算法的文本分类器的实现[D]. 电子科技大学, 2011
[16] 杨海.SVM 核参数优化研究与应用[D]. 浙江大学, 2014
[17] 巩知乐,张德贤,胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真, 2009,26(7) :164-167
[18] 王影. 基于最近邻子空间搜索的文本分类算法[D]. 北京工业大学,2014
[19] 丁勇,秦晓明,何寒晖. 支持向量机的参数优化及其文本分类中的应用[J]. 计算机仿真, 2010, 27(11): 187-190
[20] 詹增荣,曾青松. 基于径向基函数插值与 SVM 的协同过滤算法[J]. 计算机与现代化, 2015, (08): 98-103
[21] 薛松. 基于机器学习的文本处理技术研究与应用[D]. 北京邮电大学,2015
[22] 朱莹莹,尹传环,牟少敏. 一种改进的局部支持向量机算法[J]. 计算机工程与科学, 2013, 35(2): 91-95
[23] 唐守忠,齐建东. 一种结合关键词与共现词对的向量空间模型[J]. 计算机工程与科学, 2014, 36(5): 971-976
[24] 许钰. 基于半监督 SVM 主动学习的文本分类算法研究[D]. 兰州交通大学, 2013
[25] 任倚天. 基于支持向量机的海量文本分类并行化技术研究[D]. 北京理工大学, 2016
[26] 杨海. 基于改进萤火虫算法的 SVM 核参数选取[J]. 计算机应用与件, 2015,32 (6): 258-258+287
[27] K.Gayathri , A. Marimuthu Text Document Pre-Processing with the KNN for Classification Using the SVM, Proceedings of 7'h International Conference on Intelligent Systems and Control (ISCO 2013) ,IEEE, 2012:453-457

万方数据

[28] Sumedha Aurangabadkar, M.A.Potey. Support Vector Machine Based Classification System for Classification of Sport Articles. International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) . IEEE, 2014:146-150
[29] Bin Xu; Yufeng Zhang.A newSVMChinesetextofclassificationalgorithm based on the semantic kernel,2011 International Conference on Multimedia Technology, 2011:2857 – 2860
[30] Chang Liu, Zhongqiang Gao, Weihua Zhao, A New Path Planning Method Based on Firefly Algorithm,2012 Fifth International Joint Conference on Computational Sciences and Optimization, IEEE, 2012:775-778
[31] M.K.A. Ariyaratne, T.G.I. Fernando, S. Weerakoon. A Modified Firefly Algorithm to solve Univariate Nonlinear Equations with Complex Roots, International Conference on Advances in ICT for Emerging Regions (TCTer), 2015: 160-167
[32] Sankalap Arora, Satvir Singh. Mutated Firefly Algorithm, International Conference on Parallel, Distributed and Grid Computing, IEEE, 2014:33-38
[33] [33]袁锋,陈守强,刘弘,等. 一种改进的文化萤火虫算法[J]. 计算机仿真, 2014, 31(6): 261-265, 286
[34] 朱书伟,周治平,张道文. 基于改进多目标萤火虫算法的模糊聚类[J]. 计算机应用, 2015, 35(3): 685-690
[35] 李瑞青. 改进的萤火虫算法及应用[D]. 吉林大学, 2015
[36] 王振武,孙佳骏,尹成峰. 改进粒子群算法优化的支持向量机及其应用[J]. 哈尔滨工程大学学报, 2016,37(12) :1-6
[37] 陈健飞,蒋刚,杨剑锋. 改进 ABC-SVM 的参数优化及应用[J]. 机械设计与制造,2016,01 :24-28
[38] 王超学,张涛,马春森. 改进SVM-KNN的不平衡数据分类[J]. 计算机工程与应用, 2016,52(04): 51-55+103
[39] 张进,丁胜,李波. 改进的基于粒子群优化的支持向量机特征选择和参数联合优化算法[J]. 计算机应用, 2016,05 :1330-1335
[40] 冯晓琳,宁芊,雷印杰,陈思羽. 基于改进型人工鱼群算法的支持向量机参数优化[J]. 计算机测量与控制,2016,24(05):237-241
[41] 赵宇,陈锐,刘蔚. 集成特征选择的最优化支持向量机分类器模型研究[J]. 计算机科学,2016,08: 177-182+215
[42] Milan Tuba, Nebojsa Bacanin. UpgradedFireflyAlgorithmfor Portfolio Optimization Problem,2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation 2014: 11-118
[43] LiZhou Feng; WanLi Zuo; YouWei Wang.Improved Comprehensive Measurement Feature Selection Method forTextCategorization. 2015 International Conference on Network and Information Systems for Computers. 2015: 125-128
[44] Inoshika Dilrukshi,Kasun De Zoysa.Twitter newsclassification: Theoretical and practical comparison ofSVMagainst Naive Bayes algorithms. 2013 International Conference on Advances in ICT for Emerging Regions (ICTer). 2013: 278-278
[45] Aditya M. Deshpande,Gaurav Mohan Phatnani,Anand J. Kulkarni. Constraint handling in Firefly Algorithm,2013 IEEE International Conference on Cybernetics (CYBCO) 2013: 186-190.
[46] He Jia-jing, Zhang Heng-wei.Fuzzy subspace clusteringalgorithmbased on modified firefly algorithm,Third International Conference on Cyberspace

万方数据

Technology (CCT 2015), 2015: 1-6.
[47] 靳小波. 基于机器学习算法的文本分类系统[D]. 西北工业大学, 2005
[48] 付平. 人工萤火虫算法的参数分析与改进及其应用[D]. 华东交通大学, 2013
[49] 王蕾. 一种人工萤火虫群优化算法改进的研究[D]. 青岛理工大学, 2015
[50] 张明,张树群,雷兆宜. 改进的萤火虫算法在神经网络中的应用[J]. 计算机工程与应用, 2015, 12(12): 1-5

万方数据

攻读硕士学位期间发表的论文及其它成果

发表的学术论文：

[1]第二作者.中文文本分类关键技术的研究[J].电脑编程技巧与维护, 2016.14:14-15+33.

参加科研情况：

[1]北京电信光交箱实时管理系统

[2]广东电信光纤保护在线管理系统

万方数据

致谢

两年半的研究生生活一瞬而逝，在此期间的学习生涯收获颇多。在此感谢两年半以来对我产生影响的人。

首先要感谢胡朝举老师两年半来的悉心教导；胡老师严谨细致，一丝不苟的工作学习风格给我留下了深刻的印象，我在以后的工作学习中定当以胡老师为榜样，以认真端正的态度对待每一件事。同时要感谢两年半来教导过我的华电老师们，是他们让我学习更进一步。

在此，还要感谢我实验室的同学们，感谢他们在学习和生活中对我的帮助，在科研工作中给予我极大的帮助，在每位同门的帮助下，在科研和论文写作上有了极大的进步。

感谢室友两年半来对我的帮助，她们热爱生活，勤奋好学的态度都对我产生了深远的影响。

最后，特别感谢父母二十多年来的养育之恩，谢谢他们默默的付出，得以让我顺利的完成学业，以后的日子里定当回报。

谢谢我生命中的每个人。

万方数据