学校代码：10289
分类号：TP391
密　级：公 开
学　号：092070054

江苏科技大学
硕士学位论文

基于聚类算法和支持向量机算法的文本分类算法研究

研究生姓名 刘文　　导师姓名 吴 陈

申请学位类别 工学硕士　　学位授予单位 江苏科技大学

学科专业 计算机应用技术　　论文提交日期 2012 年 3 月 10 日

研究方向 计算智能与技术　　论文答辩日期 2012 年 3 月 17 日

答辩委员会主席 高 尚　　评 阅 人

2012 年 3 月 19

基于聚类算法和支持向量机算法的文本分类算法研究
刘 文
江苏科技大学

分类号： TP391

密　级： 公开

学　号： 092070054

工学 硕士学位论文

基于聚类算法和支持向量机算法的文本分类算法研究

学生姓名　刘　文

指导教师　吴陈教授

江苏科技大学

二〇一二年 三 月

A Thesis Submitted in Fulfillment of the Requirements

for the Degree of Master of Engineering

Study of Text Classification Algorithm Base on Clustering Algorithm and Support Vector Machine Algorithm

Submitted by

Liu Wen

Supervised by

Professor Wu Chen

Jiangsu University of Science and Technology

March, 2012

江苏科技大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

年　　月　　日

江苏科技大学学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、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，允许论文被查阅和借阅。本人授权江苏科技大学可以将本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和汇编本学位论文。

本学位论文属于：

(1)保密□，在___年解密后适用本授权书。

(2)不保密□。

学位论文作者签名：　　指导教师签名：

年　月　日　　年　月　日

摘 要

随着因特网的迅速发展，人们能获得的文本信息也急剧增长。如何快速的提取有效的信息是信息处理领域研究的重要内容，而文本分类是快速有效获取文本信息的主要方法，在文本分类过程中文本分类算法是保证分类速度和效果的关键技术之一，因此对文本分类算法的研究具有重要的意义，也是本文研究主要内容。

本文首先对国内外文本分类研究现状进行了详细分析，并分析了文本分词、文本特征提取、文本表示等技术。其次对常用的聚类算法和分类算法进行了详细的研究，并重点对 K-近邻算法和支持向量机算法在文本分类中的应用进行了研究。论文主要工作如下：

第一，在深入研究了 K-近邻算法基础上，针对 K-近邻算法在文本分类过程中存在的类倾斜、存储与计算量大等问题，本文提出了支持向量数据描述（support vector data description，简称 SVDD）和改进 K-近邻算法结合的分类策略。该方法首先采用 SVDD 方法对训练文本集中的各类进行裁剪，并形成新的训练文本集。然后通过类别标准差判断是否仍然倾斜，如果倾斜则对发生倾斜的类别进行收缩，形成调节因子。并通过调节因子对传统的 K-近邻判别函数进行改进。通过实验证明，本文提出的新方法能有效的解决传统 K-近邻方法的类倾斜问题，并且新方法的查全率、查准率、F1 值高于传统的 K-近邻方法。

第二，详细研究了多类分类支持向量机在文本分类中的应用，为解决传统的一对多支持向量机存在样本不平衡性和不可分区域，本文提出相应的解决方法。该方法首先采用 K-均值算法对训练集进行聚类，对每个类中不能正确聚类的文本采用一对多方法训练两类分类器，即训练对应类别的分类器，然后将训练集通过一对多 SVM 产生的分类器进行测试，将落在不可分区域的样本采用一对一方法进行再次训练，从而达到训练样本平衡和缩小不可分区域的目的。最后通过实验证明新方法在文本分类效果上优于传统的一对多支持向量机分类方法。

本文对用于文本分类的主要分类算法进行了研究，并对 K-近邻算法和支持向量机算法进行了改进，改进后的方法明显的改善了文本分类的效果，并为进一步的文本分类研究打下了基础。

关键字 文本分类；K-近邻算法；K-均值方法；支持向量机

Abstract

With the rapid development of internet technology, the amount of text information is in rapid increase. How to effectively extract these information is most important in information processing, which the main approach now is through text classification. The way to design an effective classification algorithm in text classification process is the key technique to ensure fast speed and excellent result. So it is of great significance to have research on the text classification algorithms, and text classification is the main content in this paper.

After analyzing domestic and international researches on text classification, the method of word segmentation, text feature extraction and text presentation are introduced. Then common clustering algorithms and classification algorithms are analyzed in detail. Among them, the k-nearest neighbor and support vector machine algorithm are mainly focused. The main work is listed as follows:

First, with deep research on k-nearest neighbor algorithm, a new method to deal with k-nearest neighbor classification boundary problem is proposed in this paper. Firstly, the new data set was got by cutting the training set using support vector data description algorithm. Then, standard deviation function is used to judge whether the new training set is still in imbalance. If the imbalance still exists, the shrinkage factor is brought in to shrink the class. And by shrinkage factor we improve the decision function of k-nearest neighbor. Experiments show that the method proposed can effectively solve the boundary problem in k-nearest neighbor text classification, and has higher recall, precision and F1 value.

Second, after a detailed study of multi-class support vector machine in text classification, a new method to solve the imbalance and dead zone in one multi-class support vector machine is proposed. The method has two advantages, minimizing the region in which data cannot be classified correctly in one-versus-rest support vector machine, and solving the imbalance of samples. Firstly, k-means method is used to cluster the training set. Secondly, for each text which hasn’t been clustered correctly, the one-versus-rest method was used to generate two types of classification classifier. Then, the dataset which cannot be classified by one-versus-rest will be trained again using one-versus-one method, this will reduce region of cannot classified and keep the samples in balance. Experiments show that the new method is more effective than traditional one-versus-rest method.

The main algorithms for text classification are studied in this paper, and then the

improved k-nearest neighbor algorithm and support vector machine algorithm are given. Experiments show that the new methods effectively improved the classification results, and have laid a good foundation for further study.

Keywords text classification; k-nearest neighbor algorithm; k-means method; support vector machine

目 录

摘 要 I
Abstract III
第 1 章 绪论 1
1.1 研究背景及意义 1
1.2 国内外研究现状 2
1.3 课题研究的主要内容和创新点 4
1.4 论文的组织结构 5
第 2 章 文本分类的相关技术 6
2.1 文本预处理相关技术 6
2.1.1 文本的分词技术 6
2.1.2 文本的特征选择 7
2.1.3 文本的表示 9
2.2 常用文本聚类算法介绍 10
2.2.1 K-均值聚类方法 11
2.2.2 CURE 算法 11
2.2.3 DBSCAN 聚类算法 12
2.2.4 EM 算法 13
2.3 常用文本分类算法介绍 13
2.3.1 朴素贝叶斯方法 13
2.3.2 K-近邻算法 14
2.3.3 支持向量机分类算法 14
2.3.4 神经网络算法 17
2.3.5 决策树算法 19
2.4 文本分类的性能指标 19
2.5 本章小结 20
第 3 章 SVDD 与改进的 K-近邻算法结合的分类策略 21
3.1 SVDD 基本原理 21
3.2 K-近邻文本分类算法分析 22
3.2.1 K-近邻算法存在的不足 22
3.2.2 已存在的解决方法 23

3.3 SVDD 算法与改进的 K-近邻算法结合的分类策略 24
3.3.1 解决边界问题的新方法分析 24
3.3.2 SVDD 与改进的 K-近邻结合的算法描述 26
3.4 实验结果及分析 26
3.5 本章小结 28
第 4 章 基于 K-均值聚类算法改进的多类 SVM 方法 29
4.1 常用的多类 SVM 方法 29
4.1.1 一对多 SVM 分类器 29
4.1.2 一对一 SVM 分类器 30
4.1.3 有向无环图多类 SVM 30
4.1.4 二叉树多类 SVM 31
4.2 用于文本分类的一对多 SVM 算法分析 33
4.3 基于 K-均值算法改进的一对多 SVM 方法 33
4.3.1 改进方法的原理和目标 33
4.3.2 Region_Recombine_class 算法流程 35
4.4 实验及结果分析 36
4.4.1 libsvm 介绍及使用 36
4.4.2 实验结果分析 39
4.5 本章小结 42
第 5 章 总结和展望 43
总结 43
下一步的工作 44
参考文献 45
攻读学位期间发表的学术论文目录 48
致谢 49

Contents

Abstract in Chinese I
Abstract in English III
Chaper1 Exordium 1
1.1 Research Background and Significance 1
1.2 Reasearch Stuatus 2
1.3 The main Work 4
1.4 This Organizational Structure 5
Chapter2 TheTechnologies of Text Classification 6
2.1 Text Preprocessiong Technology 6
2.1.1 The Word Segmentation Technology of Text 6
2.1.2 Text Feature Selection 7
2.1.3 Text Represention 9
2.2 General Clustering Algorithm 10
2.2.1 K-means Algorithm 11
2.2.2 CURE Algorithm 11
2.2.3 DBSCAN Algorithm 12
2.2.4 EM Algorithm 13
2.3 General Classificationg Algorithm 13
2.3.1 Na ïve Bayesian Method 13
2.3.2 K-Nearest Neighbor Alogrithm 14
2.3.3 Support Vector Machine Algorithm 14
2.3.4 Neural Network Algorithm 17
2.3.5 Decision Tree Algorithm 19
2.4 Evaluation for Text Classification 19
2.5 Chapter 20
Chapter3 Classification Strategy Base on SVDD and Improved KNN 21
3.1 Basic Principle of SVDD 21
3.2 KNN Algorithm Analysis 22
3.2.1 Disadvantages of KNN Algorithm 22
3.2.2 Existing Solution to The Shortage of KNN 23

3.3 Classification Strategy Base on SVDD and Improved KNN 24
3.3.1 The New Method of Analysis 24
3.3.2 Description of The New Method 26
3.4 Experiment and Result Analysis 26
3.5 Chapter Summary 28
Chapter4 Improvement of One-Versus-Rest SVM on K-means 29
4.1 General Multi-Class SVM 29
4.1.1 One-Versus-Rest SVM 29
4.1.2 One-Versus-One SVM 30
4.1.3 Driected Acyclic Graph SVM 30
4.1.4 Binary Tree SVM 31
4.2 One-Versus-Rest SVM Analysis 33
4.3 Improvement of One-Versus-Rest SVM on K-means 33
4.3.1 The Principle and Goal of New Method 33
4.3.2 The Flow OF Region_Recombine_class Method 35
4.4 Experiment and Result Analysis 36
4.4.1 The Introduction and Use of Libsvm 36
4.4.2 Results Analysis 39
4.5 Chapter Summary 42
Chapter5 Summary and Outlook 43
Summary 43
Outlook 44
References 45
Papers Published 48
Acknowledgements 49

第 1 章 绪论

1.1 研究背景及意义

上个世纪 80 年代以来，信息化的浪潮席卷全球，信息技术迅速地渗透到各个领域。信息的来源也扩展成多方面的，比如报纸、电视、广播等等。近十年来，随着 Internet 的普及和网络技术的不断完善，Internet 已经成为了全球最庞大最丰富的信息资源库。1998 年的统计结果表明，Internet 上有约 3.5 亿个静态 HTML 页面，每天增加将近 100 万[1]，同时，仅美国国内就有近 140 万种图书付印，这一数据还以平均每年 6 万种的速度上升[2]。2011 年 7 月 19 日，China Internet Network Information Center（简称 CNNIC）发布的《第 28 次中国互联网络发展状况统计报告》显示，截至 2011 年 6 月底 ，中国网民规模达到 4.85 亿人，比 2010 年 6 月增加 6500 万人，普及率攀升至 36.2%，比 2010 年 6 月提高 4.4 个百分点，与 2010 年 12 月相比，半年增长率为 9.4%，手机网民规模达 3.13 亿[3]，网络新闻、博客/个人空间、电子邮件、微博、网络文学、论坛等等的半年增长率在 0.1%-208.9%之间。由此可见，进入信息时代后，不仅信息的数量成爆炸式增长，而且信息的形式更加复杂，从而导致了 Internet 上信息的杂乱性和冗余性。

面对如此浩瀚并且持续急剧膨胀的信息海洋，怎样对这些信息进行有效地管理和组织，如何能更快、更准确、更全面地从如此巨大的信息库中找到用户所需要的信息是当前信息检索领域研究的重要课题。尽管网络信息数量急剧增长，表现形式也多种多样，但信息仍然以文本形式为主。这是因为文本是人们传递信息、互相学习和交流的主要载体，而其它形式的信息、如视频、图片等等都可以用文本进行标注。所以对文本信息的处理，是有效的管理和组织庞大信息资源的重要工作之一。在对大量文本数据进行处理和组织的相关技术中，文档的分类和聚类技术可以很大程度上改善信息杂乱现象，从而方便用户准确地找到所需要的信息，并且还可以分流信息。因此，对文本的自动分类的研究是一项非常有实用价值的重要技术，也得到了越来越多的研究人员的广泛关注。

文本分类的任务是根据文本的内容自动的把它分到预先定义好的类别中。文本分类技术是通过分析待分类文本，提取待分类文本的特征，比较待分类文本和系统已定义类别对象的特征，将待分类对象划分到特征最相近的一类，并赋予相应的分类号。中文文本的自动分类技术包括多种技术，其中主要包含文本的表示和文本的分类技术。

文本的表示是文本自动分类的基础，它主要将文本表示成计算机能够识别和计算的数据类型。其中主要包括分词技术、特征提取和特征权重、文本的表示模型等。

文本分类技术主要是指分类算法，它是文本自动分类的关键。20 世纪 90 年代以前，基于知识工程的分类方法是占主导地位的文本分类方法，这种方法是由专业人员采用手工进行分类。采用人工分类方法不仅费时费力，而且效率低下，无法满足人们的要求。90 年代以来，在文本的自动分类中应用较多的是机器学习方法和统计方法。目前对英文的自动分类已经取得了许多成果，提出了多种相对成熟的分类方法，如贝叶斯算法、K-近邻算法、DT 方法以及基于 VSM、SVM、回归模型等的算法。目前国内对中文文本分类的研究主要集中在 K-近邻算法、朴素贝叶斯(Naive Bayes，简称 NB)算法、决策树算法、支持向量机等技术上。

由于对中文文本的处理与分类相对比较复杂，并且起步相对较晚，也没有形成统一的横量标准，所以各个算法在不同的领域具有不同的效果，但随着信息化的不断发展，对文本分类的要求也越来越高，因此应继续研究提高文本分类的效果的方法，使中文文本分类在信息过滤、机器翻译、信息检索、信息组织和管理、邮件过滤、自动文摘等诸多领域得到更深入广泛得应用，因此对中文文本分类算法的研究具有重要的理论和实用意义。

1.2 国内外研究现状

国外对文本分类的研究相对较早，开始于二十世纪五十年代，IBM 公司的 H.P.Luhndui 在文本分类领域进行了开创性的研究[4]，他首先提出将词频统计思想用于文本分类中，并取得了很好的成果。 到上世纪六十年代，Maron 在发表的论文《On relevance，probabilistic indexing and information rendeval》中[5]，第一次提出了采用关键词进行自动分类的技术。上世纪七十年代初 Salton 等人在统计学方法基础上提出了向量空间模型，由于该模型简明地实现了对文本特性的抽象描述，从而成为文本分类处理的一种经典模型[6]。随后许多学者在这一领域进行了大量的研究工作，并取得了卓有成效的成果。在上世纪八十年代之前，基于知识工程的方法是最有效的文本分类系统，这种方法是基于人工实现的，需要相关领域的专家，还需要知识工程师编制大量的推理工作，其中卡内基集团为路透社开发的 Construe 系统是典型的代表[7]。

到上世纪九十年代，随着机器学习、数据挖掘、模式识别等技术的不断发展，新型的机器学习方法逐渐取代了基于知识工程的方法，也成为文本分类的主流技术。这种分类技术节省了大量的人力资源，加快了文本分类系统的建立速度。到目前为止，研究者们提出多种分类算法和分类模型，其中比较常见的有：贝叶斯算法、K-近邻算法、神经网络、支持向量机等等，这些算在实际中得到了很好的应用。

总之，国外的文本分类技术发展相对比较早，在文本的表示模型、文本的特征选择、文本的分类算法、文本语料库、分类算法的评价标准等方面发展已相对比较成熟。

相对于国外的文本分类发展水平，国内的文本分类发展比较落后，起步也比较晚，1981 年，南京农业大学的侯汉清教授介绍了国外在自动分类、分类检索、管理分类表、编制分类表等方面的概况[8]，并深入探讨了计算机在文本分类工作中的应用。国内对中文文本分类的研究是建立在英文文本分类基础上进行的，结合中文文本的特征，将相关技术加以改进应用于中文文本分类上，从而形成了中文文本自动分类研究体系[9]。许多学者在基于知识和统计两种方法上对中文文本分类进行了大量的研究工作，主要有基于词典的自动分类系统和基于专家系统的分类系统。1998 年底，文本数据库挖掘成为我国国家重点基础研究发展规划首批实施项目中的重要内容。此后，我国陆续研制出了一批具有代表性的中文文本自动分类系统。

近十年来，随着对文本分类研究的不断深入，取得了大量研究成果。2005 年，李荣陆等人使用最大熵模型进行了中文文本分类，并通过实验比较和分析了基于最大熵模型的分类器的分类性能[10]。姚力群和陶卿结合局部线性和单类别的思想对科技文本分类问题进行了研究，结合局部线性的思想来寻找文本样本的内在支撑流形，采用单类别的思想确定正负样本的分界面，并由此得出的对科技文本进行分类的分类方法具有可控制的正负样本分类精度，分类效果好、简便的参数估计等优点，从而提出了一条有效的解决科技文献分类问题的途径[11]。2006 年，尚文倩、陶卿等人对文本的特征选择方法做出了研究，将基尼指数应用到特征选择算法当中，并构造出适合文本分类的基于基尼指数的特征选择评估函数[12]。陈晓云、陈祎等人对现有的在文本分类中应用关联分类的不足进行了改进，将词频引入到分类规则树中，不仅可以提高关联分类的准确率，而且在不影响分类质量的情况下加快了分类速度，从而很好的弥补了传统的关联分类的不足[13]。苏金树、张博锋等人对以机器学习为基础的文本分类技术的应用进行了研究，特别指出在互联网信息进行处理等的应用中所面临的挑战，对文本分类技术的研究进展分别从算法、模型和评测等方面进行综述评论，并认为目前文本分类的关键问题主要有：数据集的偏斜、多层分类、非线性、标注瓶颈、网页分类，算法的扩展性等[14]。2007 年，王强、关毅等人提出了一种降低分类器数目提高分类精度的类别噪声裁剪算法，算法主要是对文本关键特征中蕴含的类别信息进行分析，从而能够通过类别信息主动预测待分类文本可能的类别集[15]。唐华和曾碧卿提出了一种基于遗传算法和信息熵的文本分类规则抽取算法，该算法的目的是在数据集中寻找分类规则;首先利用信息熵生成初始种群，然后利用优化的遗传算法抽取相应规则。最后在六个标准的公共领域的数据集上进行了实验，实验结果表明，所提出的算法能大大提高对知识的理解力[16]。2008 年，朱靖波等人提出了一种基于分类错误分布的混淆类识别技术[17]。李文波等人对传统的 LDA 模型进行了改进，将文本的类别信息加入 LDA 模型中，从而很好的克服了 LDA 模型强制分配隐含主题的不足[18]。2009 年，郝秀兰等人通过定义临界点对训练集的性质进行讨论，并给出了计算临界点的上下近似

值的算法。然后，根据临界点的上下近似值结合训练集样本数对传统的 K-近邻算法的决策函数进行修改，形成新的 K-近邻文本分类算法[19]。

目前，国内在中文文本自动分类领域中已经取得了令人瞩目的研究成果，其中清华大学、上海交通大学、哈工大、中国科学院等科研院所在文本分类领域做了很多的研究，一些已经被成功的推广和应用，典型的代表系统有北大天网和百度搜索等。但对于中文文本分类仍然存在很多问题，如大部分分类算法只适用于特定的领域、分类算法不完善、分类算法准确率不高以及没有统一完备的数据集等等问题。因此，文本分类的实际应用和它自身固有的特性给机器学习、人工智能提出了新的挑战，这使得文本分类的研究特别是分类算法的研究，仍然是信息处理领域重要的研究课题之一。

1.3 课题研究的主要内容和创新点

根据上节对国内外研究现状的分析可以了解到，国内外对文本的特征选择、表示方法、分类算法等都进行了相关的研究，并取得了一定的成果。但对于文本分类仍然存许多问题，特别是分类算法仍需要做进一步研究和改进。本文首先详细介绍了文本分类的相关技术，然后主要针对文本分类中的算法进行研究。其中详细介绍 K-最近邻算法和支持向量机算法在中文文本分类中的应用，并具体的分析他们的优缺点。在此基础上对支持向量机算法和 K-近邻算法进行了改进，从而寻找一种分类效果更好的分类算法。

主要创新点如下：

（1）K-近邻算法是一种应用比较广泛且简单、有效、无参数的分类方法，但在分类过程中存储所有的训练样本，直到分类时才进行计算，并且存在类倾斜现象影响分类精度。针对 K-近邻算法存在的缺点，本文提出基于支持向量数据描述的改进的 K-近邻算法，首先利用支持向量数据描述算法的分类特点对训练文本集进行裁剪形成新的训练文本集，并根据训练集中各类的样本分布特点对 K-近邻算法的判别函数引调节因子，从而很好的解决 K-近邻算法存在的类倾斜现象，提高了分类准确率。

（2）支持向量机在多类分类时构造的一对多分类器，虽然可用于大规模文本分类，但是当文本类别数过多时，某一类的文本数目将远远少于其他剩余样本的总数，这种文本数的不平衡将影响分类的准确性，且生成的分类器会形成不可分区域。针对一对多分类器存在的缺点，本文结合 K-均值算法、一对多和一对一支持向量机算法的特点，提出新的分类算法，改进了单纯用一对多支持向量机算法时存在的样本不平衡问题，并减小了一对多支持向量机的不可分区，达到更好的分类效果。

1.4 论文的组织结构

本文共五章，各章的内容组织如下：

第 1 章，本章首先介绍了中文文本分类的研究背景和意义，然后分析了文本分类国内外的研究现状，介绍了本文主要研究工作。最后给出本文的整体组织结构。

第 2 章，本章首先介绍了文本聚类和分类的定义，然后分别介绍了中文文本的表示、中文文本的权重计算和特征提取、文本聚类算法和文本分类算法等内容，最后介绍了文本分类的评价标准。

第 3 章，本章首先分析了 K-近邻算法的优缺点，及详细介绍了支持向量数据描述算法的基本原理，然后给出了对 K-近邻算法的改进，最后介绍了实验及和结果分析。

第 4 章，本章主要介绍了多类支持向量机的基本原理、多类分类器的构造、结合 K-均值算法对多类 SVM 分类算法的改进、以及实验分析。

第 5 章，本章对全文的研究工作进行简要总结，并介绍了今后研究工作的重点，以及对研究工作的展望。

第 2 章 文本分类的相关技术

在文本信息处理中，对文本的分类和聚类是重要的处理技术，文本分类是指对文本集按照一定的分类体系或标准进行自动分类。分类的目的是建立一个分类函数或分类模型，该模型能将某一文本映射到模型中的某一个类别中。文本聚类是指根据文本具有的特征将文本集合分成多个类别或者簇的过程，聚类的结果是使同一类别中的文本具有较高的相似度，不同类中的文本内容差别较大[20]。文本聚类是典型的无指导的学习方法。

文本聚类、文本分类是两个相互关联，但又属于不同的领域的研究课题，把未知文本根据训练集合分配到已知的类别当中的过程是文本分类的主要工作。而文本聚类没有训练集合，它是根据给定文本内在的联系将文本分为不同的簇，并给每个簇一个标识。虽然文本聚类和文本分类属于不同的领域，但都需要将文本表示成电脑可以理解的形式，也就是需要对文本进行预处理。本章将在下面分别介绍文本预处理相关技术、常用聚类方法、常用分类方法及分类评价等内容。

2.1 文本预处理相关技术

将文本转化为机器能识别的形式的过程叫做文本预处理过程，中文文本的字与字、词与词之间没有特殊的分隔符，所以对中文进行预处理比英文文本复杂的多。一般文本预处理过程包括：分词，特征选择、权重计算及模型表示等。

2.1.1 文本的分词技术

对文本的预处理首先要进行文本的分词，分词是指将连续的语句切分成词序列的过程。目前常用的分词方法主要有三种：字符串匹配法、语义法和基于统计的方法。

字符串匹配法又叫做机械分词方法，它是按照一定的步骤将待分词的汉字串与一个机器词典中的词条进行匹配，若在词典中找到相应的字符串，则匹配成功[21]。按照扫描方向的不同，机械分词方法可以分为正向匹配和逆向匹配。按照不同长度优先匹配的情况，可以分为最大匹配和最小匹配；字符串匹配方法是比较成熟常用的方法，但他的分词精度及速度受机器词典和匹配方法影响较大，且不能很好的从语义上切分词语，因此在采用此方法时需要使用完善的词典及合适的匹配算法。

基于语义的分词方法是通过让计算机模拟人对句子的理解[22]，在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。这种分词方法是近几年研究的热点，由于汉语语法及句法复杂多变，相同的词在不同的语法环境中表达不同

的词义，因此通过上下文的语义对语句进行分词是十分有意义的。但这种分词方法相对不成熟，没有形成完善的机制，目前只在特定的领域和特定的问题中得到应用。

基于统计的方法是指对语句中相邻出现的各个字的组合的频度进行统计，计算他们的相互信息作为依据进行分词。此方法的依据是，文本的语句从形式上是由字与字连接而成，在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词[23]。这种方法只需对语料中的字组的频度进行统计，不需要切分词典，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些出现频度高、但并不是词的常用字组，例如“这一”、“无的”、“你的”、“很少的”等，并且对常用词的识别精度差、时空开销大。实际应用的统计分词系统都要使用一部基本的分词词典进行串匹配分词，同时使用统计方法识别一些新的词，即将字频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

2.1.2 文本的特征选择

经过分词、去除停用词等操作后，文本集的特征项个数是巨大的，往往达到几十万、几百万个，而在向量空间表示中每一个特征值就代表一维，对具有如此巨大维数空间的文本集进行分类效率是很低的，甚至是无法进行的。因此必须降低文本集的维数空间，以此提高分类速度和分类精度。

在文本的所有特征中，各个特征项对文本的分类作用是不同的，有些特征能很好的体现文本的类别，而有些特征项与文本类别关系不大，因此选择能很好的表示文本类别的特征项，是降低文本维数的重要手段。随着文本分类技术的不断进步和完善，特征选择算法已相对成熟，常见的文本特征选择算法有 DF（Document Frequency，文本频率）、MI（Mutual Information，互信息）、IG（Information Gain，信息增益）、 CHI 统计、ECE(Expected Cross Entropy，期望交叉熵)、几率比等。

（1）DF 方法　DF 是指根据包含某个特征项的文本数目来判断此特征项的重要程度。在特征选择的过程中普遍认为，如果训练集中包含某特征项的文本数目很少，则说明该特征词对类别的标示作用很小。因此文本频率方法就是通过设定一个阈值，当某文本特征的 DF 值小于这个阈值，则删除这些特征项。这种方法具有对训练文本规模的线性计算复杂度，因此容易用于大规模样本集。

DF 方法是一种简单的特征选择方法，在实际应用中也有很好的效果，但大量的实践也证明部分稀有词可能在某些类文档中含量比较多，也含有大量的类别信息，一味的删除这些词将影响分类效果。

（2）MI 方法 MI 方法主要思想是：假设类 A 中文本含有某个特征词比较多，而这些特征词在其他类别的文档中出现的比较少，则认为这些特征词与类 A 的互信息比

较大，与其他类得互信息比较小。MI 公式为：

$$MI(w,c_i)=\log\frac{p(w/c_i)}{p(w)} \tag{2.1}$$

也可以表示为：

$$MI(w,c_i)\approx\log\frac{p\times N}{(p+c)\times(p+q)} \tag{2.2}$$

其中，w 为特征项，c_i 表示类别，p 表示在类 c_i 中包含 w 的文本数，q 表示除 c_i 类之外的类别中包含 w 的文本数，c 为 c_i 类中其他特征项的文本频率之和，N 为训练文本总数。

（3）IG 方法　IG 方法是通过特征项出现前后对文本的分类结果的影响程度，来表示此特征的重要程度，计算公式如下：

$$IG(w)=p(w)p(c_i/w)\log\frac{p(c_i/w)}{p(c_i)}+p(\bar{w})\sum_i p(c_i/\bar{w})\log\frac{p(c_i/\bar{w})}{p(c_i)} \tag{2.3}$$

其中，w 表示特征项，$p(c_i/w)$ 和 $p(c_i/\bar{w})$ 分别表示当文本中出现 w 和不出现 w 时属于类 c_i 的概率，$p(c_i)$ 为 c_i 出现在文本集中的概率，$p(w)$ 为特征项 w 在文本集中出现的概率，$p(\bar{w})$ 为特征项 w 不在文本集中出现的概率。

（4）CHI 方法　CHI 方法又叫 χ^2 统计法，该方法是假设特征项与类别之间符合具有一阶自由度的 χ^2 分布，特征项与类别之间的 χ^2 值越大，表示特征项包含该类的特征信息就越多。特征项 t 和类 c_i 的 CHI 公式为：

$$CHI(t,c_i)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{2.4}$$

其中，N 为文本集包含的文本数，A 为属于 c_i 且包含特征 t 的文本数，B 为属于 c_i 且不包含特征 t 的文本数，C 为不属于 c_i 且包含特征 t 的文本数，D 为不属于 c_i 且不包含特征 t 的文本数。

（5）ECE 方法　ECE 方法也是基于概率的方法，与 IG 方法类似，不过该方法只考虑特征项出现在文本中的情况，计算公式为：

$$ECE(w)=p(w)\sum_i p(c_i/w)\log\frac{p(c_i/w)}{p(c_i)} \tag{2.5}$$

其中，$p(w)$ 为特征项 w 在文本集中出现的概率，$p(c_i/w)$ 为当文本中出现 w 特征时属于类 c_i 的概率，$p(c_i)$ 为 c_i 出现在文本集中的概率。

ECE 反映的是文本类别的概率分布和出现了某个特征词的条件下文本类别的概率分布之间的距离。$ECE(w)$ 的值越大，表示特征项 w 对文本类别分布的影响越大。

（6）几率比（Odds Ratio）方法 几率比方法比较适合于两类分类器，它只关心目标类别而不将所有类等同看待，计算公式为：

$$OR(w) = \log \frac{P(w/c_i)(1-P(w/\bar{c_i}))}{p(w/\bar{c_i})(1-p(t/c_i))} \tag{2.6}$$

其中，$P(w/c_i)$表示文本包含特征项 w 且属于c_i类的概率，$P(w/\bar{c_i})$表示包含 w 且不属于c_i类的概率。

以上各种文本特征提取算法各有优缺点，文献[24]中对 DF、IG、MI、CHI 四种方法进行实验，实验表明 IG 方法优于他三种方法，文献[25]中对 DF、IG、MI、CHI、期望交叉熵、几率比等方法进行实验，实验表明扩展的多类几率比方法分类效果最好，其次是 IG 和 ECE，但在数据均匀分布时 CHI 效果比 IG 和 ECE 要好。文献[26]中对 MI、CHI、ECE 等进行实验，实验表明 MI 效果是最好的。从文献[24-26]的研究看出，各种特征提取方法的效果与训练集的性质及分类算法密切相关，没有那个方法具有绝对的优势。因此在选择特征选择算法时要根据具体情况决定。

2.1.3 文本的表示

计算机不具有像人类大脑一样的理解力，可以通过阅读文章，结合自己的知识判别出文章的类别。因此在文本分类过程中，如何将无结构化的文本表示成计算机识别的结构化的形式，是文本分类研究的重要内容之一。

通常文本是由字、词和词组组成，将组成文本的字、词或词组叫做文本的特征项，因此可以简单的将文本表示为特征项的集合，虽然这将丢失大量关于文本内容的信息，但这样可以将文本形式化表示；即给定一篇文本 D，可以将其表示为$D = \{t_1, \ldots, t_n\}$，其中 t_i (i=1….n)表示文本的特征项。目前常见的特征表示模型有布尔逻辑模型、向量空间模型、概率推理模型等。其中应用最广效果相对较好的是向量空间模型。

（1）布尔模型 布尔模型是简单严格的文本表示模型，对文本中的特征权重采用二值化的方法来定义，即如果文本中出现特征项 t_i，则 t_i 对应的权重为 1，否则为 0。

（2）向量空间模型 向量空间模型是由 Salton[27]等人提出，最早在 smart 系统中得到应用，他主要是将文本采用向量形式表示，即由文本的特征项和特征项的权重组成的向量来表示文本，其中，特征权重表示特征项对文本类别的重要程度。因此给定文本 D，则 D 的向量空间表示为：$D = D(t_1, w_1; \ldots; t_n, w_n)$；可以简记为 $D = D(w_1, \ldots, w_n)$，其中 t_i 为文本的特征项，w_i 为 t_i 对应的特征权重，i=1…n。

衡量两个文本之间的相似程度是通过文本之间的相似度来判断的，向量空间表示模型中常用的相似度公式有向量内积和向量夹角余弦两种表示方法。给定两个文本 $D_1 = D_1(w_{11}, \ldots, w_{1n})$，$D_2 = D_2(w_{21}, \ldots, w_{2n})$，则向量内积公式和向量夹角余弦公式分别为：

$$sim(D_1, D_2) = \sum_{i=1}^{n} w_{1i} * w_{2i} \tag{2.7}$$

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{n} w_{1i} * w_{2i}}{\sqrt{(\sum_{i=1}^{n} w_{1i}^2)(\sum_{i=1}^{n} w_{2i}^2)}} \tag{2.8}$$

其中公式 2.7 中 w 表示文本特征项的权重，传统的特征权重表示为：如果特征项存在，则特征向量该特征维上的值为 1，不存在则值为 0；这样计算虽然简单，但太粗糙，不能体现该特征项在文本中的重要程度，因此这种计算方法慢慢的被代替掉。

（3）概率推理模型 概率推理模型也是比较常用的文本表示模型，它主要将用户的兴趣与文本按照概率方式进行融合，它全面考虑特征词频、文本频率和文本长度等因素。对于给定文本 d 和用户兴趣 h 的相关公式为：

$$Sim(d, h) = \sum \log \frac{p_i(1-q_i)}{q_i(1-p_i)} \tag{2.9}$$

其中，$p_i=A_i/A$, $q_i=(C_i-A_i)/(C-A)$。C 表示训练文本集的文本数，C_i 表示训练文本集中包含 t_i 特征项的文本数，A 表示训练集中与用户兴趣相关的文本数，A_i 表示在 A 个相关文本中包含 t_i 的文本数目。

以上三种文本表示模型各有优缺点，在解决不同问题中有不同的效果，在文本表示中除了表示模型之外特征权重的计算也是非常重要的，随着文本挖掘及人工智能研究的深入，涌现出了许多计算权重方法，其中 TF*IDF 是用的最广最多计算方法，他的基本思想是：特征项在同一文档中出现的频率越高越重要，在越多的文档中出现越不重要，TF 表示特征项 D 在某一文档中出现的频率，$IDF=log（N/n），N$ 表示文档总数，n 表示包含特征项 D 的文档数量，则完整的 TF*IDF 公式为：

$$W(t_i, D_j) = \frac{(1+\log(TF_{ij}))(\log(N/n_i))}{\sqrt{\sum_k [(1+\log(TF_{ik})(\log(N/n_k))]^2}} \tag{2.10}$$

其中 $W(t_i,D_j)$表示特征项 t_i 在文档 D_j 中的权重，TF_{ij} 表示特征项 t_i 在文档 D_j 中的次数。

2.2 常用文本聚类算法介绍

目前聚类算法有很多，主要可以划分为以下几种:

1.基于划分的方法：划分算法将数据集合分割为一个平坦结构的分区，根据划分数目创建一个初始划分，根据某种度量方法不断地迭代改进划分结果。

2.层次的方法：层次的方法创建给定数据对象集的层次分解。层次之间的关系表示类之间的关系。

3.基于密度的方法：此方法根据样本邻域内样本的分布密度设定阈值，当样本邻域内其他样本超出阈值则继续聚类。

4.基于网格的方法：基于网格的方法是把对象空间量化为有限数目的单元，形成一个网络结构。所有的聚类操作都是在这个网络结构（量化空间）上进行的。

5.基于模型的方法：基于模型的方法为每个簇集假定一个模型，并寻找对给定模型的最佳拟合。

下面具体介绍文本分类中常用的聚类方法。

2.2.1 K-均值聚类方法

K-均值聚类算法是基于划分的聚类方法[28]，通过给定初始聚类中心不断地移动样本形成新的簇，然后用每个簇的所有样本的平均值作为新的聚类中心，不断地迭代计算聚类中心直到聚类中心不再发生变化或满足某个约定为止，K-均值的具体算法描述如下：

（1）输入含有 N 个样本的集合；

（2）根据 K 值随机选取 K 个数据作为初始聚类中心，形成 K 个簇；

（3）计算每个样本到各聚类中心的相似度，根据相似度的大小，将样本合并到相应的簇中形成新的 K 个簇；

（4）计算每个簇中所有样本的平均值，以平均值作为每个簇的聚类中心；

（5）迭代执行（3），（4）两步，直到某个准则函数不在明显变化或者聚类的对象不再变化为止；

K-均值算法对于大型数据是可伸缩的和高效的，但 K-均值方法也存在如下缺点：第一，他受初始聚类中心影响很大，即选择不同的初始聚类中心会有不同的聚类结果；第二，初始聚类 K 值的选择，在聚类过程中是无法事先确定 K 值的，从而聚类的效果影响很大。

2.2.2 CURE 算法

CURE(Clustering Use Representatives)算法是基于层次的聚类方法[29]，它对传统聚类方法中类的表示方法进行了改进，传统的方法用所有样本或简单用聚类中心和半径等单一条件来表示一个类，而 CURE 方法是从每个类中抽取固定数量、分布较好的点作为描述此类的代表点，并引入一个适当的收缩因子，每一个代表点乘以收缩因子，从而使它们更靠近类的中心点位置。在聚类过程中每一个类用多个点来表示，这样在聚类的外延处就会向着非球形的形状进行扩展，弥补了大多数聚类算法只偏好球形和

相似大小的不足，从而那些非球形的类也可以得到表达。另外，使用收缩因子还可以降低噪音对聚类效果的影响。同时，为了提高算法的空间和时间效率，CURE 算法将随机抽样与划分相结合用于聚类过程。

CRUE 算法的主要流程如下：

（1）从原始数据中抽取一个随机样本 T。

（2）将样本 T 分为 m 个划分。

（3）将每个划分局部聚类成 k 类，其中 k>1。

（4）通过随机采样消除异常数据，若一个簇增长太慢，就删除该簇。

（5）对局部的簇进行再聚类，对于落在每个新形成的聚类中的代表点，则根据用户定义的收缩因子收缩或向簇中心移动。这些点将用于代表并描绘出聚类的边界。

（6）对簇中的数据标记上相应标记。

2.2.3 DBSCAN 聚类算法

DBSCAN(Density-Based-Spatial-Clustering of Application with Noise)算法是基于密度方法的聚类算法，该算法将具有足够高密度的区域进行聚类，并可以在带有“噪声”的数据库中发现任意形状的簇。它定义簇为密度相连的点的最大集合[30]。在介绍它的基本思想之前先介绍一下相关概念如下：

（1）对象邻域：即给定一个半径为 r 的区域。

（2）核心对象：如果一个对象的 r 邻域内，至少含有最少对象数目（Minpts），就称这个对象为核心对象。

（3）直接密度可达：在训练集合 S 中，如果样本 p 在 q 的 r 邻域内，而 q 是一个核心点，则称 p 到 q 是直接密度可达的。

（4）密度相连：p 到 s 是密度可达的，s到q是密度可达，则 p 到 q 是密度相连的。

（5）噪声： 分类结束时不包含在任何类中的样本点 。

DBSCAN 算法的核心思想是：用户给定最少样本数 MinPts 和邻域半径 r,然后通过检查数据库中的每个点的 r 邻域来进行聚类。如果一个样本 s 的 r 邻域包含的样本数多于 MinPts，则创建一个以 P 为核心对象的新簇。然后反复的寻找从这些核心对象直接密度可达的对象，把密度可达的样本合并为一个新簇，当没有新的点添加到任何簇中时聚类结束。

DBSCAN 算法的优点是可以发现任意形状的簇，并且对数据的输入顺序不敏感，具有处理异常数据的能力。它主要缺点是对用户自定义的参数非常的敏感，不同的 r 和 MinPts 对聚类的效果影响很大，导致聚类的结果差别很大，而且这两个参数带有很大的主观色彩不好确定，并且伸缩性不好，所以算法效率不是很高。

2.2.4 EM 算法

EM(Expectation Maximization)又叫期望最大化方法[31]，该算法是一种比较简单且容易实现的算法，是一种进行极大似然估计的有效的方法，它是在观察数据的基础上添加一些“隐含的数据”，从而简化计算并完成一系列简单的模拟或极大化，而不是直接对后验分布进行模拟或极大化。该算法每一次迭代都由一个极大步和期望步构成。

EM 算法的主要思想是，首先对混合模型的参数进行初始的估计，然后反复地对每个对象进行重新打分，打分的依据是根据参数向量产生的混合密度，并采用重新打分后的对象更新参数估计。通过对每个对象赋予一个概率，反映假定它是给定簇的成员时具有一定的属性值集合的可能性。

2.3 常用文本分类算法介绍

分类算法是训练文本分类器的核心，目前流行的文本分类算法主要可以归为三大类，一类是基于统计的方法，如 K-近邻算法、贝叶斯算法、类中心向量法、SVM（Support Vector Machine，支持向量机）等方法，二是基于连接的方法，如神经网络等，三是基于规则的方法，如决策树算法、粗糙集等。几类分类算法对不同领域不同的问题分类效果各有所长，下面主要介绍几种应用比较广泛，分类效果比较好的分类算法。

2.3.1 朴素贝叶斯方法

朴素贝叶斯方法是机器学习和人工智能领域应用广泛的算法之一[32]，他基于贝叶斯理论，假设事物的所有属性是相互独立互不干扰的，利用事件的先验概率、条件概率及全概率，判断对象属于各类的概率，概率值最大的类就是测试对象的类别。

在文本分类中，贝叶斯方法假设文本的所有特征是相互独立的，给定待测文本 $d_i=\{w_1,w_2,...,w_m\}$，文本集 $N=\{c_1,c_2,...,c_k\}$，则由贝叶斯公式知，给定文本 d_i 属于 c_j 的概率为：

$$p(c_j / d_i)=\frac{p(c_j)p(d_i / c_j)}{p(d_i)} \tag{2.11}$$

$$p(d_i)=\sum_{j=1}^{k} p(c_j)p(d_i / c_j) \tag{2.12}$$

$P(c_j)$ 为每个的先验概率，采用拉普拉斯概率进行估计值，公式为：

$$p(c_j)=\frac{1+N_{c_j}}{k+N_c} \tag{2.13}$$

N_{c_j} 为 c_j 中的文本数， N_c 为文本集中总得文本数，$p(d_i/c_j)$的计算公式为：

$$p(d_i / c_j) = \prod_{s=1}^{m} p(w_s / c_j) \tag{2.14}$$

d_i 的最终类别是通过 $P（c_j/d_i）$ 的值来判断，d_i 属于 $P（c_j/d_i）$ 值最大的一类。

2.3.2 K-近邻算法

K-近邻算法简称 KNN（K-Nearest Neighbor）算法,是基于统计学习理论的比较成熟的方法[33]，该方法思路比较简单清晰，具有很好的分类效果。该算法的基本思想是：寻找与待测样本近邻的 K 个样本，分别计算 K 个样本中各类的权重，权重最大的类即为待测样本的类别。在文本分类应用中，具体分类过程为：设文本向量 $w = \{t_1, t_2, \ldots, t_n\}$，其中 $t_i(i=1...n)$ 表示文本向量的特征项，训练文本集 $S = \{c_1, \ldots, c_m\}$,m 表示训练集中的类别数，其中 $c_i = \{w_1, \ldots, w_q\}$ 表示 S 中的第 i 类中的 q 个文本，i=1…m。对于待分类文本 W ,首先计算 W 与训练样本集 S 中的所有样本之间的相似程度，按照从大到小的顺序选出 K 个最相似的样本，然后统计属于 C_i 类的样本数 K_i，最后通过计算判别函数的值来确定待测文本的类别。

其中样本之间的相似程度有多重衡量标准，其中最常用的有文本距离度量法和文本相似度度量法，基于距离的度量方法常用的为欧式距离，给定两个文本 d_1=（W_{11}，W_{12}，...，W_{1m}），d_2=（W_{21}，W_{22}，...，W_{2m}），m 为文本的特征维数，

则欧式距离公式为： $D(d_1, d_2) = \sqrt{(w_{11} - w_{21})^2 + ... + (w_{1m} - w_{2m})^2}$ (2.15)

文本相似度公式为：$sim(d_1, d_2) = \dfrac{\sum_{k=1}^{m} w_{1k} \times w_{2k}}{\sqrt{\sum_{k=1}^{m} w_{1k}^2} \sqrt{\sum_{k=1}^{m} w_{2k}^2}}$ (2.16)

总体来讲 K-近邻算法是简单有效，无参数化的分类方法，该方法直接利用文本与文本之间的关系，减少了由于文本类别特征选择不当对分类效果的影响，但传统的 K-近邻算法仍存在一些缺点，如 K 值的确定不当对分类结果影响很大、对训练集的依赖性大需存储所有的训练文本等等，对 K-近邻算法的改进是本文的重点内容之一，本文将在第 3 章详细介绍。

2.3.3 支持向量机分类算法

支持向量机算法简称 SVM（Support Vector Machine）算法，该算法建立在统计学习理论中的 VC 维和结构风险最小化基础之上，并结合最优化理论来得到分类决策函数的分类算法。其基本思想是寻找一个分类超平面，将两类样本分到超平面的两侧[34]。

他在解决非线性问题、高维模式识别问题等许多问题中显示出许多优势，是统计学习理论中比较实用的算法之一，目前已在人脸识别、手写数字识别、文本分类、信息检索等领域得到成功应用。

2.3.3.1 两类线性可分 SVM

线性可分是指可以找到一个分类超平面将两类正确的分开。给定文本训练集 $T=\{(x_1,y_1),(x_1,y_1),\ldots(x_k,y_k)\}$ 是 $R^n\times Y^k$，其中，$x_i\in R^n$，$y_i\in Y=\{1,-1\}$，i=1,…k，若存在 $w\in R^n$，$b\in R$ 和正数 δ，使得集合满足：

$$\begin{cases}(w\cdot x_i)+b\geq\delta & 当y_i=1\\(w\cdot x_i)+b\leq-\delta & 当y_i=\text{-}1\end{cases}\qquad i=1,..k \tag{2.17}$$

则称文本训练集 T 是线性可分的[35]。

在线性可分的情况下 SVM 的基本思想如图 2.1 所示，图中黑点和白点是两类分类文本，L 为两类的最优分类线，L_1、L_2 是过离分类线距离最近的样本，且平行于分类线的直线，两线之间的距离叫做分类间隔，SVM 算法就是寻找一个最优分类线，即不仅能将两类分开，还能使分类间隔最大的分类线，则将公式(2.17)进一步归一化得：

$$y_i[(w\cdot x_i)+b]\geq 1\qquad i=1,...k \tag{2.18}$$

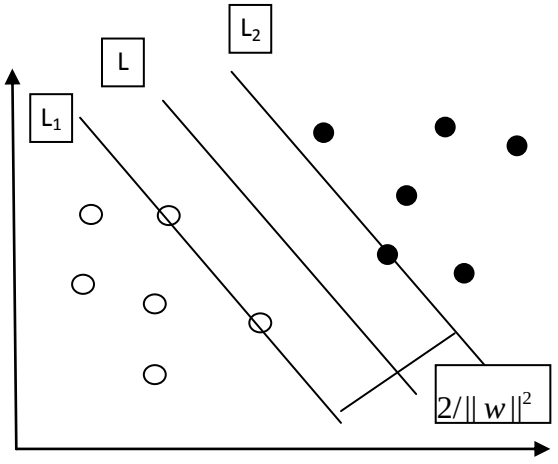

图 2.1 最优分类线

Fig.2.1 The optimal classification line

此时 L_1、L_2 分别满足方程 $y_i[(w\cdot x_i)+b]=1$ 和 $y_i[(w\cdot x_i)+b]=-1$，分类间隔距离为 $2/\|w\|^2$，最优分类面满足公式（2.18），且使 $2/\|w\|^2$ 最大，则在 L_1、L_2 上的文本就叫做支持向量，要使 $2/\|w\|^2$ 最大，相当于使 $\frac{1}{2}\|w\|^2$ 最小，这样问题就转化为一个解非

线性规划为题：

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|^2\\&s.t.\quad y_i[(w\cdot x_i)+b]\geq 1\qquad i=1,2,...,k\end{aligned}\tag{2.19}$$

上述解最优分类面问题，根据最优化理论存在唯一最小解。其 Lagrange 函数为：

$$L=\frac{1}{2}\|w\|^2+\sum_{i=1}^{k}\alpha_i[1-y_i(w\cdot x_i+b)]\qquad \alpha_i\geq 0\tag{2.20}$$

其中，α_i 为 lagrange 乘子。对上式中 w，b 求导，并令其导数为零，即：

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}\alpha_i y_i x_i=0\Rightarrow w=\sum_{i=1}^{k}\alpha_i y_i x_i\tag{2.21}$$

$$\frac{\partial L}{\partial b}=\sum_{i=1}^{k}\alpha_i y_i=0\tag{2.22}$$

将公式（2.21）、（2.22）代入（2.20）中，将求解最优问题转化为其对偶问题：

$$\max w(\alpha)=\sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)\tag{2.23}$$

$$\sum_{i=1}^{k}y_i\alpha_i=0\tag{2.24}$$

$$\alpha_i\geq 0,i=1,...,k\tag{2.25}$$

通过求解上面的对偶问题可以求出最优分类面，假设 α^* 为最优解，则只有少部分 α_i^* 不为零，其他都为零，不为零时对应的 x_i 为支持向量，$w^*=\sum_{i=1}^{k}\alpha_i^* y_i x_i$ 为训练文本的线性组合，选用任意 x_i 就能求出 $b^*=y_i-w^*\cdot x_i$，从而得到分类决策函数为：

$$f(x)=\mathrm{sgn}(w*\cdot x+b^*)=\mathrm{sgn}(\sum_{i=1}^{k}\alpha_i^* y_i x_i\cdot x+b^*)\tag{2.26}$$

判断待测文本的类别时可以通过公式（2.26）来分类。

2.3.3.2 两类线性不可分 SVM

线性不可分是指类别之间无法用一个最优分类面完全的分离开。此时可以适当放宽公式（2.20）中的限制条件，引入一个松弛因子 ξ_i，则公式（2.18）变为：

$$y_i[(w\cdot x_i)+b]\geq 1-\xi_i\qquad i=1,...k\tag{2.27}$$

当 ξ_i 足够大时所有的文本都将包括在内，为了限制 ξ_i 无限放松引入惩罚因子 C，加入限制条件：

$$c\sum_{i=1}^{k}\xi_i \quad i=1,...,k \quad c>0 \tag{2.28}$$

此时将问题转化为另一个求最优化问题，如下：

$$\min_{w,b}\frac{1}{2}\|w\|^2+c\sum_{i=1}^{k}\xi_i \tag{2.29}$$

$$s.t. \quad y_i[(w\cdot x_i)+b]\geq 1-\xi_i \tag{2.30}$$

$$\xi_i\geq 0 \quad i=1,...,k \tag{2.31}$$

转化为其对偶问题：

$$\max w(\alpha)=\sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{2.32}$$

$$\sum_{i=1}^{k}y_i\alpha_i=0 \tag{2.33}$$

$$0\leq\alpha_i\leq c, i=1,...,k \tag{2.34}$$

经过转化后的判别函数与线性可分时类似，经过转化后的解决线性不可分问题的 SVM 方法，可以防止由噪声产生引起的分类效果差问题。

以上主要介绍了基于两类文本分类问题的 SVM 方法，对于多类文本分类问题的 SVM 是本文研究重点将在后面章节介绍。

2.3.4 神经网络算法

神经网络是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法。它是由神经元组成的并行处理网络，每个神经元具有一个单一的输出联接，不过可以根据需要把这个输出联接分支成多个并行的输出联接，而且这些并行联接都输出相同的信号，即相应神经元的信号，信号的大小不因分支的多少而变化[36]。根据网络结构和学习算法的不同，人工神经网络可分多层感知器、自组织映射和 Hopfield 网络等。下面以 BP 神经网络为例说明神经网络在文本分类中的应用。

BP 神经网络是一种按误差逆传播算法训练的多层感知器网络，他具有一个输入层，一个输出层和至少一个中间层。正如图 2.2 所示，输入层各神经元负责接收输入信息，并传递给中间层各神经元；中间层是内部信息处理层，负责信息变换，根据信息变化能力的需求，中间层可以设计为单隐层结构或者多隐层结构，用来进一步处理中间层传递到输出层各神经元的信息，完成一次学习的正向传播处理过程，由输出层向外界输出信息处理结果。当实际输出与期望输出不符时，进入

误差的反向传播阶段。误差通过输出层，按误差梯度下降的方式修正各层权值，向中间层、输入层逐层反传。反复的进行信息正向传播和误差反向传播过程，就是各层权值不断调整的过程，也是神经网络学习训练的过程，此过程一直进行到输出误差减少到能接受的程度，或者达到设定的学习次数为止。

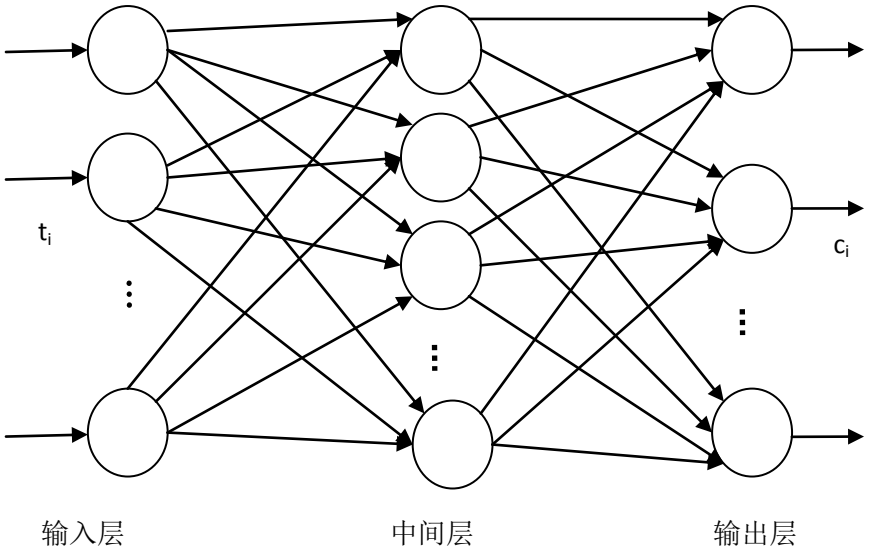

图 2.2 神经网络分类图

Fig.2.2 The use of neural networks in classes

设在三层 BP 神经网络中，输入文本向量为 $d=(t_1,t_2,...t_k)$，k 代表文本向量的特征项个数，输出类别向量 $C=(c_1,c_2,...c_h)$，h 表示类别数。那么在这个三层 BP 神经网络中，输入层为 k 个神经元，中间层为 n 个神经元，输出层为 h 个神经元。中间层的神经元个数 n 要根据具体问题而定。输入层和中间层之间、中间层和输出层之间的连接权值，在神经网络的训练阶段根据训练样本学习得到。在对训练样本的判断学习中，若输出正确，则保持权值不变，否则就对权值进行必要的调整，使神经网络判别出正确的结果。经过多次的判断学习并对权值不断调整后，神经网络最终趋于稳定，此时就可以使用此 BP 神经网络进行文本分类。

由以上分析知，在文本分类过程中输入层的第 i 维的取值为：

$$t_i=\begin{cases}1 & \text{文本中存在第i个特征项时}\\ 0 & \text{文本中不存在第i个特征时}\end{cases} \qquad i=1,...,k$$

输出层的第 j 个分量取值为：$c_i=\begin{cases}1 & \text{文本属于第j类}\\ 0 & \text{文本不属于第j类}\end{cases}$ $j=1...h$

由于神经网络是模拟人的神经系统，因此具有很强的自组织自学习能力，并且具有很好的鲁棒性，在模式识别及文本分类等领域取得了很好的成果。

2.3.5 决策树算法

决策树是基于树形结构的多级分类算法，由根节点，非叶子节点和叶子节点构成。每一个非叶子节点代表对一个或多个特征属性的测试，叶子节点代表某一个类别，从根节点到叶子节点的一条路径就是对相应对象的一条分类规则，可以将决策树很容易的转化为分类规则，是一种分类模式比较直观的分类方法。

决策树构造过程可以分为两步。第一步，决策树的生成：通过训练样本集生成决策树的过程。第二步，决策树的剪技：决策树的剪枝是对上一阶段生成的决策树进行完善的过程，主要是用测试数据集中的数据对产生的初步规则进行校验的过程，将那些影响预测准确性的分枝剪除。在构造决策树的过程中，关键的是内部节点的特征项选择，即在非叶子节点用那一个特征项对样本进行决策，通常采用信息增益来计算每个节点的特征项值，根据特征项的值大小，每次选择值最大的特征项作为当前节点的测试属性，以便降低各子集中的不同类别的混合程度。

目前在决策树推导、决策树属性选择、决策树的可扩展性等方面都有大量的研究，并开发了许多基于决策树的算法和系统，其中比较有代表性的有 ID3、ID4、C4.5、SLIQ、SPRINT 等等。

2.4 文本分类的性能指标

文本分类算法的好坏主要是看分类器的分类结果来衡量，对文本的分类结果主要从计算复杂度、描述的简洁度和有效性三个方面来评估[37]。计算复杂度分为空间和时间复杂度，如果按照分类的步骤来分，计算复杂度又可以分为训练和分类计算复杂度。描述的简洁就是算法描述的简洁程度，也可以理解为算法理解的难易程度。有效性是代表一个分类器正确分类的能力；在这三个方面中分类器的有效性最为重要，因此对有效性的评估是分类器评估工作的主要内容。

有效性的评估中查准率（Precision，简称 p）和查全率（Recall，简称 r）是在中英文文本分类中最常用的指标，查准率是指对测试文档进行分类后，真正符合分类意图的文本比例，它体现了系统检索结果的准确程度。查全率表示被检索出的结果文档集中真正符合检索意图的文档数在所有符合检索意图的文档集中所占的比率，它体现了系统检索的完备性。查准率和查全率的公式可分别表示如下：

查准率： $p = cp_i / k_i$ (2.35)

查全率：$R = cp_i / c$ (2.36)

其中 c 是实际属于 c_i 类的样本数，k_i 是分类器预测为 c_i 的样本数，cp_i 是正确分

到类 c_i 的样本数。

查准率和查全率的值越高代表分类器的性能越好，但这两个标准又是相互克制的，即单纯提高查准率就会导致查全率的降低，单纯提高查全率就会导致查准率的降低。所以，为了避免某一个指标过低，一个好的分类算法需要在这两者之间做一些折中。因此采用用 F1 值来描述查准率和查全率的综合效果，每类的 F1 值公式为：$F1 = 2P_iR_i/(R_i + P)$ (2.37)

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率；总体 *F1* 值为各类 *F1* 值的平均值，从公式可以看出 *F1* 表示了 *P* 和 *R* 的平衡关系，只有当 *P* 和 *R* 都大时 *F1* 才可能大，因此 *F1* 值综合反应了分类器的整体性。

2.5 本章小结

本章首先介绍了文本预处理技术中的分词、特征选择、文本表示等相关技术，然后对常用的文本聚类和文本分类方法做出了介绍，讲解了各算法的主要思想和重要算法步骤，最后对文本分算法的分类性能评价方法做出了讲解，讲解了目前常用的查准率、查全率和 F1 值三种评价指标定义及公式。

第 3 章 SVDD 与改进的 K-近邻算法结合的分类策略

3.1 SVDD 基本原理

只由一种类别的样本作为训练样本集的分类问题叫做单类分类问题，通常这类样本叫做正常类[38]，而其它数据叫做异常类或负类。单类分类问题的典型方法有基于密度的方法和基于边界的方法。

基于密度的方法是首先需要通过参数化或非参数化来估计样本的概率密度，然后通过设置阈值判断待测数据属不属于正常数据，但现实中的目标数据往往反映的是数据的区域，并不是密度，所以采用密度的方法很可能把正常数据的稀疏区域作为低密度区而误判，并且基于密度的估计方法要把样本映射到一个高维空间，因此该方法适合用于低维空间。

而基于边界的估计方法是根据样本的区域特征来构造一个超球面，或其他几何体，使得这个超球体或其他几何体包含样本尽量多的情况下半径尽量的小，训练过程是寻找样本的支持，所以可以用于高维、多数据和有噪声点的样本集。

由于文本分类是多数据、高维的数据集，所以采用基于边界的方法更适合文本类分，并且基于边界的方法已得到广泛的应用和研究[39]。其中 SVDD 算法是一种应用十分广泛的基于边界的单类分类算法，并基于支持向量机算法提出的另一种比较重要的核算法[40]。SVDD 算法的基本原理如下：

给定包含 n 个样本点的训练集合 $X=\{x_1,\ldots,x_n\}$，寻找一个超球体，使其在包含样本数尽可能多的情况下，而使超球体的半径尽可能的小。由于敏感的样本点往往离球心比较远，因此允许一些样本点在球体的外面，在此引入松弛变量 ξ_i，设球心为 O，半径为 R，得到下面的限制条件：

$$(x_i-o)(x_i-o)^T \le R^2+\xi_i \tag{3.1}$$

其中，$\xi_i \ge 0$，所以要使球体半径 R 和松弛变量 ξ_i 这两项最小化：

$$f(R,o,\xi_i)=R^2+c\sum_i \xi_i \tag{3.2}$$

为了平衡不被包含的样本的数目和球体的体积引入常数 c。在加入了限制条件（3.1）后构造 Lagrange 函数：

$$L(R,o,\alpha_i,\xi_i)=R^2+c\sum_i \xi_i-\sum_i \alpha_i\left\{R^2+\xi_i-(x_i-2ox_i+o^2)\right\}-\sum_i \beta_i\xi_i \tag{3.3}$$

其中 $\alpha_i \ge 0$，$\beta_i \ge 0$，则可将问题(3.2)转化为其对偶问题：

$$\min_{o} \sum_{i,j=1}^{n} o_i o_j (x_i \bullet x_j) - \sum_{i=1}^{n} o_i (x_i \bullet x_j) \tag{3.4}$$

$$s.t. \sum_{i=1}^{n} o_i = 1, \alpha = \sum_{i=1}^{n} o_i x_i \tag{3.5}$$

$$0 \le o_i \le c, i = 1, ..., n \tag{3.6}$$

通常通过比较一个样本点到圆心的距离和半径，来判断样本点属于正常类，还是非正常类，当一个测试样本点 y 的距离小于半径时，样本点 y 为正类，否则为非正常类。

$$(y-o)(y-o)^T = (y \bullet y) - 2\sum_{i=1}^{n} (y \bullet x_i) + \sum_{i,j=1}^{n} o_i o_j (x_i \bullet x_j) \le R^2 \tag{3.7}$$

能够使上面等式成立，并且满足 $O_i \neq 0$ 的向量叫做支持向量，少量球面上的支持向量决定了球体的半径。

3.2 K-近邻文本分类算法分析

K-近邻算法用于文本分类中，算法实现由训练过程和分类过程两部分构成。训练阶段需要对训练集建立特征词词典、统计特征词词频、训练分类器等。分类阶段需要将测试文本根据特征词词典进行向量化，并采用训练分类器判断待测文本的类别。K-近邻算法是一种有效的文本分类算法，在文本分类中得到广泛应用。本节将重点分析影响 K-近邻算法分类性能的原因，及相关改进方法。

3.2.1 K-近邻算法存在的不足

K-近邻算法是一种简单、有效、无参数的文本分类算法[41]，受到广大研究者的关注，并取得了一定的进展。但该算法在文本分类过程中仍存在如下不足：

（1）分类过程中存在类边缘问题，会导致测试文本类别的误判。所谓类边缘问题是指当某些类别的样本数目远比其它类别的样本数目多时，分类器在分类时往往偏向于大类，从而使原本属于小类的样本错判别为大类的现象，这种现象也叫做类倾斜现象。如图 3.1 所示，类 1 和类 2 代表两类文本，在两类的边缘处由于类 1 的样本数量明显比类 2 多，因此在圆圈中属于类 2 的样本 x 很容易被误判为类 1。

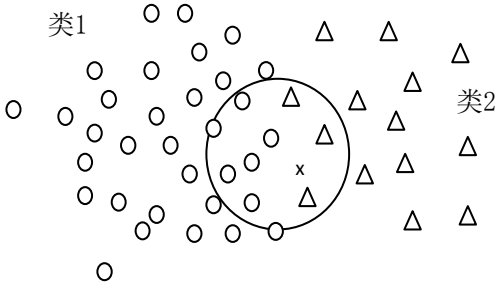

图 3.1 KNN 分类示意图

Fig.3.1 KNN classification

对于给定一个训练集如何判断该训练集是否存在类边缘问题，往往要考虑训练集各类样本的总体分布情况。由于标准差是与样本的总体分布相关的，因此可以从标准差出发来判断。具体定义如下：

定义 1：给定一个训练文本集$T=\{c_1,c_2,...,c_m\}$，m(m>2)为训练文本集中包含的类别数，其中各类的样本数分别为 N_i(i=1,…,m)，设训练集的标准差为δ,最小值为$x_0=\min(N_1,N_2,...,N_m)$。当$x_0<\delta$时，称文本训练集 T 存在类倾现象，否则不是类倾斜的。

定义 1 中的标准差公式为：$\delta=\sqrt{\frac{1}{n}\sum_{i=1}^{m}(N_i-\bar{N})^2}$ (3.8)

其中，$\bar{N}$ 为训练集的平均值。

假设一个具有 4 个类的训练集，4 个类的样本数分别为{9,9,9,81}，则标准差δ为 36，最小值 x_0 为 9，由于 $x_0<\delta$，根据定义 1 该训练集是类倾斜的。

（2）K-近邻算法是一种懒散的方法，在分类过程中 K-近邻算法必须将所有的训练样本保存，并且需要计算待测样本与所有训练样本之间的距离，造成相当大的存储和计算开销。

因此，在采用 K-近邻算法应用于文本分类时需要考虑这两方面的不足，尽量减小待测样本的错分，提高 K-近邻算法的分类效果。

3.2.2 已存在的解决方法

针对 K-近邻算法的类倾斜现象已存在多种解决方法，具体解决方法有改变度量方式、重取样和代价相关学习等方法。

（1）改变评价度量标准

该方法是指对分类结果进行评估时采用针对性的评价方法，从而减少使用分类准确率进行评估带来的误差。其中 ROC 曲线、COST 曲线、F2 度量、Laplace 估算等是比较常用的针对类倾斜现象的评估方法。

（2）重取样方法

常用的随机重取样有随机上取样、随机下取样和混合取样三种方法。对重取样相关学者做了广泛的研究，其中文献[42]中采用 SMOTE 下取样方法，文献[43]中 Jo Taeho 等人采用基于聚类的上取样方法，文献[44]中 Batista、Prati 等人采用 SMOTE 和 TOMEK 结合的混合取样方法，以及文献[45]中提出的 DataBoost 采用合成数的取样方法。

随机取样虽然一定程度上解决了类倾斜缺陷，但取样方法也存在一些缺陷，如随机下取样有可能移除重要的样本,而随机上取样存在会产生过度拟合问题。

（3）代价相关学习

代价相关学习方法的主要思想是指正确识别稀有类的价值要远远超过正确识别普通类的价值，并且假设在分类过程中对于不同类型错误的代价矩阵是已知的。然而，这个代价矩阵往往是未知的，而且代价相关学习与重取样是紧密联系的。通常来说代价相关学习往往要比随机重取样方法好，但精心设计的重取样及组合方法要比代价相关学习更好。

3.3 SVDD 与改进的 KNN 算法结合的分类策略

3.3.1 解决边界问题的新方法分析

由上节分析可知，K-近邻方法在文本分类过程中存在存储问题和类倾斜现象，而这些不足是由算法原理和训练集中各类文本的数目差异过大造成的。因此结合 K-近邻算法的分类特点，本文从训练集各类的文本裁剪和 K-近邻判别函数两方面出发来解决类倾斜问题。

1. 各类的文本裁剪

根据近邻原则，当待测文本处于类中心区域时，无论类中心区域有没有训练文本，只要边界区域有足够的文本便可以确定待测文本的类别，当待测文本处于类边界区域时，他的近邻绝大部分也存在于边界区域，也就是处于类中心区域的文本对分类是不起作用的[46]。因此可以对训练文本集中各类中心区域的文本进行裁剪，由于 SVDD 对单类问题具有良好性能，并且通过构造一个超球体来进行分类，可以很好的区分中心区域文本和边缘区域文本，因此本文利用 SVDD 对训练集中的各类文本进行裁剪。

2. K-近邻判别函数的改进

对一个有 m 个类的训练集传统的 K-近邻的判别函数为：

$$f(w) = \arg\max f_i(w) \tag{3.9}$$

$$f_i(w) = \sum_{w_i \in knn(w)} sim(w, w_j) d(w_j, c_i) \tag{3.10}$$

其中，i=1,..,m，j=1,…,K，w 为待测文本，$knn(w)$ 表示 w 的 k 个近邻组成的集合，

$sim(w,w_j)$表示文本 w 与 w_j 的相似度， $d(w_j,c_i)$ 表示 w_j 是否属于 c_i 类， $d(w_j,c_i)$的值为：

$$d(w_j,c_i)=\begin{cases}1 & w_j\in c_i\\ 0 & w_j\notin c_i\end{cases} \tag{3.11}$$

由上面的判别函数可已看出，传统的判别函数没有考虑训练文本数量信息，在一定 K 值下分类结果向大类倾斜。因此本文从文本分布出发，引入调节因子 $g(\beta)$ 对判别函数按各类文本数目进行指数运算的加权处理，即令 $g(\beta)=1/N_i^{\beta}$，对判别函数（3.9）进行调节，使判别函数适当的往小类倾斜。

要想求解调节因子 $g(\beta)$，首先需要计算 β 的值，给定一个 m 类的训练集 T,各类的文本数目为 $N_1,N_2,...,N_m$，当 T 是类倾斜时，首先引入收缩因子 β，得到一组不倾斜的文本数目 $N_1^{\beta},N_2^{\beta},...,N_m^{\beta}$，即满足 $\mathrm{x}_0=\min(N_1^{\beta},...N_m^{\beta})\geq\delta'$，当 $\mathrm{x}_0=\delta'$ 时为临界点，因此可以通过临界点求出 β，即：

$$\mathrm{minN}_i^{\beta}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(N_i^{\beta}-\frac{\sum_i^m N_i^{\beta}}{m}\right)^2} \tag{3.12}$$

由公式（3.12）可以看出，当 β 值为 0 时等式右边为 0，当 β 值为 1 时，等式左右两边为原训练集数值，因此以上两种取值是没有意义的，所以 β 的取值范围为（0,1）。

由公式（3.12）直接求解 β 是很困难的，因此本文采用迭代法求解 β 的值，令 β 初值为 0.99，步长为 0.01，具体求解步骤如下：

$k=0.99$ ；

do

{

$x_i=(x_0)^k$ ；

$N_i'=(N_i)^k\ (\mathrm{i}=1,\ldots,\mathrm{m})$ ；

$\delta_i=stdev(N_1',...,N_m')$ ；

$k=k-0.01$ ；

}

$\mathrm{while}(x_i>\delta_i)$

$\beta=k$ ；

其中 N_i 为第 i 类的文本数目，$\mathrm{x}_0=\min(N_1^{\beta},...N_m^{\beta})$，$k$ 为 β 的初始值，根据以上方法在一定的误差范围内可以求的收缩因子 β，得到一组不倾斜的样本数目 $N_1^{\beta},N_2^{\beta},...,N_m^{\beta}$，从而得到调节因子 $g(\beta)=1/N_i^{\beta}$ 的值。从而得到改进的判别函数为：

$$f(w)=\begin{cases}\arg\max f_i(w) & \text{当不存在倾斜时}\\ \arg\max[f_i(w)g(\beta)] & \text{当存在倾斜时}\end{cases} \tag{3.13}$$

3.3.2 SVDD 与改进的 K-近邻算法结合的算法描述

针对 K-近邻算法的不足，根据 3.3.1 节对新算法的分析，将 SVDD 与改进的 K-近邻算法结合的分类算法核心思想描述如下：

（1）首先对训练集中每类文本采用 SVDD 进行裁剪，对训练集中每类样本采用 SVDD 求出各类的超球面，由于类中心区域的文本并不都是无用的，过量的裁剪将会影响分类效果，所以不能一味的裁剪中心区域的文本，为此引入区域半径 μ（$0<\mu<R$）作为样本裁剪的区域半径，其中 R 为超球面半径，选取每类中到圆心 O 的距离大于 μ 的文本为新的训练文本。

（2）其次，根据新生成的训练集是否倾斜采用不同的 K-近邻判别函数，即若新的训练文本集仍是类倾斜的，则采用改进的判别函数，若不是倾斜的则采用原判别函数。分类器总体训练流程如图 3.2 所示。

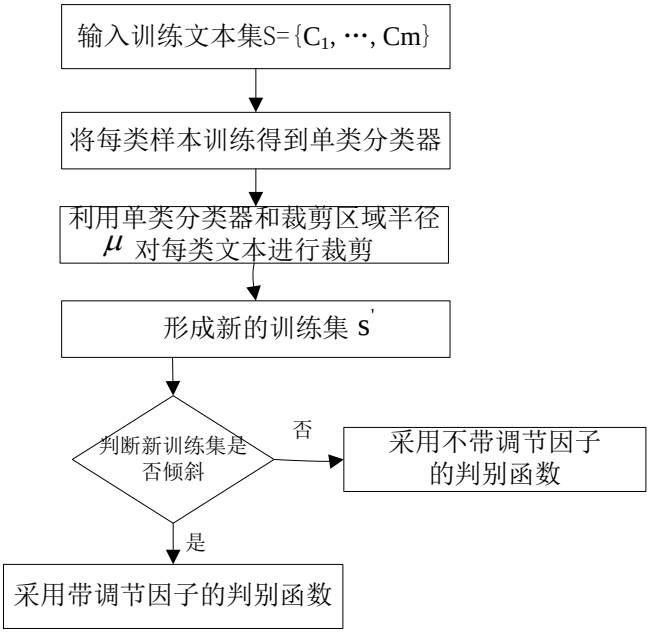

图 3.2 分类器训练流程图

Fig.3.2 process of training classifier

3.4 实验结果及分析

实验数据采用中科院计算所提供的 TanCorpV1.0 中文文本语料库,该语料库共 12 大类 14150 篇文本。该语料库提供的文本形式分两种，一种为原始文本，即没有处理过的文本，一种是分完词的形式。由于本文研究重点为分类算法，所以本文采用处理过的文本形式，这样可以大大减少本文的工作量。本实验从中抽取了体育、财经、电脑、科技、房产等五类文本，为了验证本文的方法，选取训练样本集时使各类文本数

目相差很大，其中各类别文本数为 1200、170、740、800、200。测试样本集各类文本数都为 100，即本实验原始训练集文本数为 3110 篇，测试集为 500 篇。本文的实验在 MATLAB 环境下进行。

实验共分两个阶段进行：

第一阶段主要验证裁剪区域半径的取值对样本裁剪率的影响，即选取合适的裁剪半径 μ，试验中选取 μ 值分别为 R/8,R/7,R/6,R/5,R/4,R/3,R/2,2R/3 进行裁剪，裁剪结果如图 3.3 所示，其中横坐标为 μ 值，纵坐标为各类剩余文本数目。

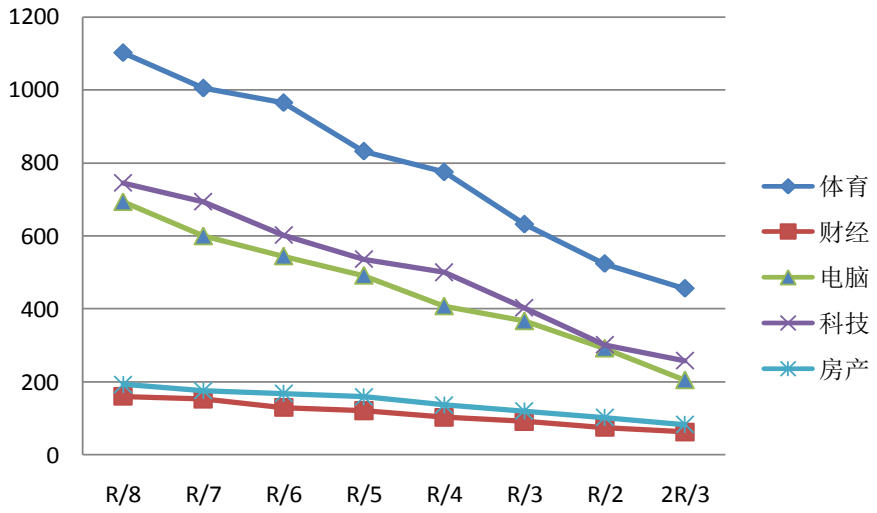

图 3.3 裁剪结果

Fig.3.3 cut out text of results

从图 3.3 可以看出随着裁剪半径的增加被裁减掉的样本也随之增加，如果裁剪半径选取过大将会丢失大量的文本信息，因此结合文本分类的经验，本实验选取裁剪半径 μ 的值为 R/3.

第二阶段主要进行分类器训练，及文本测试。经过裁剪后新文本集各类文本数目为 632，92，367，403，120，通过计算可知训练集仍是倾斜的。本实验主要与已存在的重取样方法进行比较，其中传统的 K-近邻算法作为比较的基准线。在文献[43,47]中提到，在很多情况下重取样的最佳取样率为 0.8，因此实验中取 0.8 为上取样的增加率和下取样的减少率。

实验结果采用查全率、查准率、F1 值等国际评价标准来评价分类效果，对以上实验数据，分别用传统的 K-近邻算法、SVDD 与改进的 K-近邻结合的算法、随机上取样和随机下取样算法等进行实验。各算法的实验结果如表 3.1 所示。其中 KNN 表示传统的 K-近邻算法，SVDD β-KNN 表示采用 SVDD 与改进的 K-近邻结合的算法，upsamp 表示随机上取样算法，downsamp 表示随机下取样算法。

表 3.1 实验结果比较

Table 3.1 The result of the experiment

类名	评价标准	KNN	SVDD β-KNN	upsamp	downsamp
体育类	查全率	0.7500	0.7900	0.7800	0.7700
	查准率	0.7426	0.7980	0.7959	0.8370
	F1	0.7463	0.7940	0.7879	0.8021
财经	查全率	0.6200	0.7400	0.7100	0.7000
	查准率	0.6078	0.7253	0.7245	0.7071
	F1	0.6138	0.7326	0.7172	0.7035
电脑	查全率	0.7300	0.8100	0.7700	0.8000
	查准率	0.7684	0.8438	0.7938	0.8333
	F1	0.7487	0.8266	0.7817	0.8263
科技	查全率	0.7200	0.7900	0.7800	0.7600
	查准率	0.7346	0.8061	0.7960	0.7677
	F1	0.7272	0.7980	0.7879	0.7638
房产	查全率	0.6400	0.7500	0.7000	0.7200
	查准率	0.6667	0.7653	0.6931	0.7273
	F1	0.6531	0.7576	0.6965	0.7187

由表 3.1 可以看出，由于财经和房产的训练样本数相对较少，受其他类别样本影响较大，采用传统的 K-近邻算法时查全率、查准率、F1 值都较低，而 upsamp、downsamp 效果差不多，且都明显高于传统的 KNN 算法，而本文的 SVDD β-KNN 算法的查全率、查准率、F1 值明显高于 upsamp 和 downsamp 算法。

从总体上看，对各类的分类效果 SVDD β-KNN 好于 upsamp 和 downsamp 算法，upsamp 和 downsamp 分类效果差不多，且都好于传统的 KNN 算法。

实验证明，当存在类倾斜现象时，本文提出的算法明显改善了传统的 K-近邻算法的分类效果，在查全率，查准率及 F1 明显升高，缓解了小类测试文档被误判为大类文档的现象，充分证明了本文算法的有效性。

3.5 本章小结

本章首先介绍了单类分类问题的基本概念及 SVDD 算法的基本原理，分析了传统的 K-近邻算法的不足，然后详细的分析了本文提出的 SVDD 算法与改进的 K-近邻算法相结合的原理及算法流程，最后通过实验验证了本章提出算法的有效性。

第 4 章 基于 K-均值聚类算法改进的多类 SVM 方法

在第 2 章中已经介绍了用于两类文本分类的 SVM 方法的基本原理，但现实生活中文本的分类往往是多类问题，为了使多类 SVM 在文本分类中得到更好的应用，本文在综合分析研究多类 SVM 处理文本分类的基础上，提出了一种基于 K-均值算法的多类 SVM 文本分类方法。

4.1 常用的多类 SVM 方法

采用 SVM 对多类文本进行分类时，多类分类 SVM 的实现主要有两种思想：

第一种是直接解决多类问题的思想，在原来构造两类 SVM 的算法基础上，构造具有多个值的分类模型，对新模型的目标函数进行最优化处理，求出一个分类判别函数，可以通过这个判别函数判断出所有类别[48]。但采用这种方法的目标函数非常复杂不易求解，并且效率也不高，所以这种方法很少被采用。

第二种方法是采用两类 SVM 分类器来实现，通过两类 SVM 构造方法构造多个两类分类器，最后分类时结合所有的两类分类器的结果判断测试样本的类别。这种方法比第一种方法容易实现，所以得到广泛的研究。

下面介绍以第二种方法实现的常用的多类 SVM 方法。

4.1.1 一对多 SVM 分类器

一对多（one-versus-rest，OVR）SVM 分类器是最早用于多类分类的 SVM 算法[49]，主要思想是，对于输入的 K 类训练样本，训练生成 K 个两类分类器，第 i 个分类器的训练样本是这样确定的，即将第 i 类的所有样本作为正类，其余类的所有样本作为负类。

由一对多 SVM 分类器的构造原理可知，给定训练集 $(x_1,y_1),\ldots,(x_n,y_n)$，$y_i \in \{1,...,k\}$ i=1,…,n，y_i 为 x_i 的类别标号，第 i 个分类器是求解如下问题：

$$\begin{aligned} &\min_{w^i,b^i,\xi^i} \frac{1}{2}\| w^i \|^2 + c^i \sum_{j=1}^{n} \xi_j^i \\ &w^i \cdot \phi(x_j) + b^i \geq 1 - \xi_j^i, \text{当} y_j = i \\ &w^i \cdot \phi(x_j) + b^i \leq -1 - \xi_j^i, \text{当} y_j \neq i \\ &\xi_j^i \geq 0, \quad j = 1,...,n \end{aligned} \tag{4.1}$$

求解 K 个这样问题，便得到 K 个决策函数：

$$\begin{aligned} &f_1(x)=\mathrm{sgn}(w^1\cdot\phi(x))+b^1 \\ &..... \\ &f_1(x)=\mathrm{sgn}(w^k\cdot\phi(x))+b^k \end{aligned} \tag{4.2}$$

当测试样本时，将测试样本分别通过(4.2)中的每个判别函数，具有最大函数值的分类器对应的类别为该样本的类别。函数表达式如下：

$$class(x)=\arg\max_{i=1,\ldots,k}(f_i(x)) \tag{4.3}$$

4.1.2 一对一 SVM 分类器

一对一 SVM（One-Versus-One SVM，OVO SVM）分类器是通过在每两类之间训练一个两类分类器，即给定有 K 个类的训练集，需要训练$K*(K-1)/2$个两类分类器。例如第 i 类和第 j 类的样本作为训练集时，令第 i 类标记为正类，第 j 类标记为负类，则 i 类和 j 类的优化问题为：

$$\min_{w^{ij},b^{ij},\xi^{ij}}\frac{1}{2}\|w^{ij}\|^2+c^{ij}\sum_{t=1}^{n}\xi_t^{ij} \tag{4.4}$$

$$w^{ij}\cdot\phi(x_t)+b^{ij}\geq 1-\xi_t^{ij} \qquad 当 y_t=i \tag{4.5}$$

$$w^{ij}\cdot\phi(x_t)+b^{ij}\leq -1+\xi_t^{ij} \qquad 当 y_t=j \tag{4.6}$$

其中，$\xi_t^{ij}\geq 0 \quad t=1,\ldots,n$

对于$k*(k-1)/2$个这样的优化问题，可以得到$k*(k-1)/2$个决策函数：

$$f_{ij}(x)=\mathrm{sgn}(w^{ij}\cdot\phi(x)+b^{ij}) \quad \text{i,j=1,…k 且 } i\neq j \tag{4.7}$$

对待测样本进行分类时，一对一 SVM 算法采用投票机制，通过每一个$f_{ij}(x)$判别函数进行测试，并对相应的类别投一票，最后测试完成后得票最多的类别就是测试样本的类别。

4.1.3 有向无环图多类 SVM

有向无环图多类 SVM（Directed Acyclic Graph Support Vector Machines，DAGSVM）是基于一对多方法上构造的分类策略，有向无环图是指图中每一条边都有方向但没有环的图[51]。

有向无环图多类 SVM 分类方法在训练过程中和一对一 SVM 分类方法一样，即如果训练集具有 k 个类别，那么需要训练$k*(k-1)/2$个分类器，所不同的是在测试阶段，有向无环图多类 SVM 分类方法是通过构造一个 k 层的有向无环图来实现分类，其中有向无环图由内部节点和叶子节点构成，内部节点为$k*(k-1)/2$个分类判别函数，叶

子节点为 k 个类别标示。因此有向无环图方法的主要测试过程为：首先将测试文本经过第一层的根节点判断，根据根节点的判断结果，进入下一层判断，如此循环操作直到到达叶子节点为止，叶子节点代表的类别即为该样本的类别。

图 4.1 是由有 4 个类别的训练集构成的有向无环图，其中 4 个类的类别标示“1”、“2”、“3”、“4”为叶子节点，表示分类结果；内部节点由 6 个决策函数构成，每一个内部节点都与下一层分类决策函数相连。从图可以看到，第一层由类别 1 和类别 4 生成的分类器构成，如果结果为“非 1”，即不属于第一类则进入由 2 和 4 构成的分类器，如果结果为“非 4”，则进入由类别 1 和类别 3 构成的分类器中进行判断，如此循环直到到达叶子节点为止，叶子节点代表的类别就是该样本的类别。

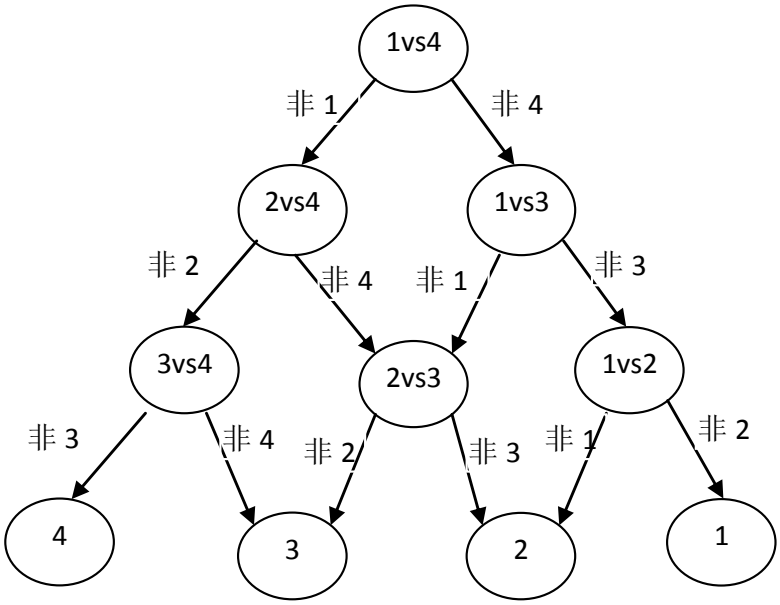

图 4.1 有 4 个类别的 DAG SVM 分类示意图

Fig.4.1 The description DAG SVM for four class

4.1.4 二叉树多类 SVM

二叉树多类 SVM 方法也是基于两类分类器构造而成[52]，并且一个具有 M 个类别的训练集需要训练 M-1 个分类器，非叶子节点为分类器，叶子节点为各类的类别标示。

构造一棵具有 M 个叶子节点的二叉树的方案有多种，对具体问题可以根据以下情况选择二叉树 SVM 的构造方案：

（1）在对 M 类中的所有数据基本没有先验知识的情况下，则无法指导叶子节点的划分，因此可以采用每次决策即可以分出一类的二叉树结构。

这种二叉树的 M-1 个分类器训练过程为：训练第 1 个分类器，则第 1 类的所有样本为正类，第 2,…,M 类的样本为负类样本；训练第 i 个分类器，则第 i 类的所有样本为正类，第 i+1,…,M 类的样本为负类样本；直到第 M-1 个分类器，则第 M-1 类的所

有样本为正类，第 M 类的样本为负类样本。

采用上述方法构造的具有 4 类样本的示例如图 4.2 所示，一个待测样本最多只需要 3 次即可以判断出属于那一类。

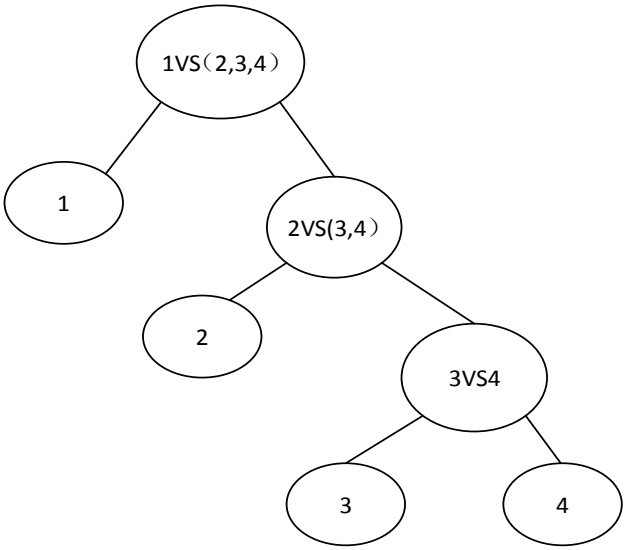

图 4.2 二叉树 SVM

Fig.4.2 the description of binary tree SVM for four classes

（2）如果对 M 类训练样本具有部分先验知识，例如对体检人员进行分类，可以知道体检人员可以包括男性老年人，女性老年人，男童，女童等，但不清楚那一类多，那一类少，所以可以采用完全二叉树形式。

这种形式的训练过程是将所有训练集样本分成两子类，再将子类分成两个子类，如此循环直到得到一个单独的类别为止，一个具有 8 类的训练集生成的二叉树如图 4.3 所示，

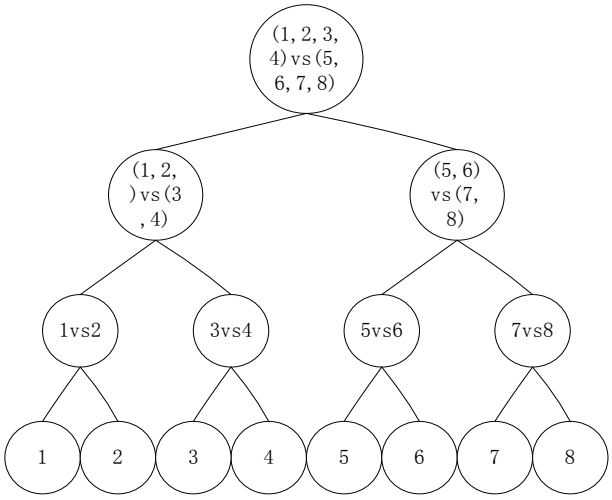

图 4.3 完全二叉树 SVM

Fig.4.3 the description of binary tree SVM for eight classes

(3)如果对训练集中每类的样本有足够的先验知识，则可以根据先验知识结合二叉

树理论构造一棵测试速度最优的二叉树。

4.2 用于文本分类的一对多 SVM 算法分析

将一对多 SVM 用于文本分类，整个实现过程分为训练和分类过个过程，具体过程为：给定一个 K 个类别的文本训练集，首先进行训练分类器，根据一对多 SVM 原理，训练构造 K 个两类分类决策函数，其次对测试文本进行分类，对测试文本 x 分别经过训练过程中产生的 K 个分类决策函数，计算各个决策函数的值，值最大的决策函数对应的类别就是该文本的类别。

通过分析一对多 SVM 的分类器训练过程，及文本分类过程可知，一对多 SVM 用于多类文本分类时与其他多类 SVM 方法相比具有简单直观、构造两类分类器个数少、决策速度快等优点。但该方法也存在如下缺陷：

（1）在生成每一个两类分类器时，训练集所有的文本都需要参加计算，并且每个分类器在训练的过程中正类的文本数量往往远远小于负类的文本数量，这种在分类器训练过程中的样本不平衡性，不但增加了计算复杂度还影响分类准确率。

（2)在训练过程中生成的分类器存在不可分区域，以具有三个类别的训练集为例，采用一对多 SVM 时的不可分区域如图 4.4 阴影所示，图中 f1、f2、f3 分别是对应类 1、类 2、类 3 形成的分类面，可以看出当测试文本处于阴影部分时将不属于任何类别。

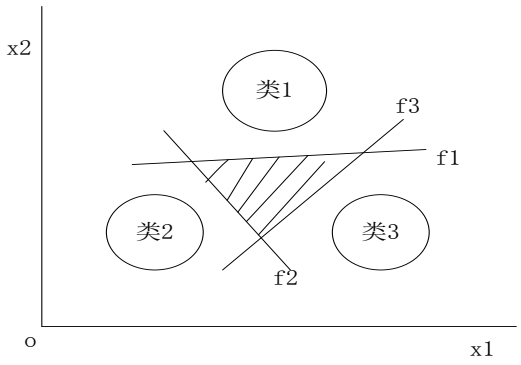

图 4.4 一对多方法的不可分区域

Fig.4.4 Unclassifiable regions by the OVR-SVM

由于一对多 SVM 存在以上缺点，因此在采用一对多 SVM 方法用于文本分类时需要考虑以上两点对分类效果的影响。

4.3 基于 K-均值算法改进的一对多 SVM 方法

4.3.1 改进方法的原理和目标

通过上节的分析可知，一对多 SVM 在训练过程中存在文本不平衡的缺点，并且

存在不可分区域。因此本文针对一对多 SVM 方法存在的不足分别采取相应的措施进行改进，以达到改进传统一对多 SVM 方法的目的，提高传统一对多 SVM 分类算法在多类文本分类中的分类效果。具体改进如下：

1.解决分类器生成过程中训练文本数目的不平衡性

根据 4.2 节分析可知，一对多方法的文本不平衡性是指每次作为正类的文本和作为负类文本数量相差很大，从而影响分类器的分类效果。又由 SVM 的基本原理可知，训练两类分类器就是寻找两类的最优分类线，而最优分类线的确定是由少量支持向量文本决定的，对于支持向量大都分布在各类的边缘处，也就是对于训练分类器来说大部分文本是无用的，只有少量边界上的支持向量是有用的。因此可通过减少作为负类的各类中的无用文本，从而可以达到正负类文本数量相对平衡的目的。

由聚类的基本原理可以知道，聚类是将具有不同属性或相似度的文本聚成不同簇的过程，有第二章相关介绍可知，K-均值是一种简单常用的聚类算法，而且 K-均值是基于聚类中心不断迭代的聚类方法，即各类中距离聚类中心位置近的文本将被划为一类，而各类的边缘和交叉处的文本往往将会被错分，而这些被错误聚类的文本区域也是支持向量文本所在的区域，因此支持向量机一般包含于被错聚类的文本之中。

根据以上分析，本文首先采用 K-均值算法对文本训练集的各类进行聚类，在传统的 K-均值聚类过程中初始聚类中心的选择是随机的，没有先验知识的，但对文本训练集聚类时，是知道各类文本的类别标示的，因此可以选取合适的初始聚类中心，如选择各类文本的平均值为初始聚类中心，从而可以降低初始聚类中心对聚类结果的影响，并且使的聚类生成的簇标示与训练集中各类的类别一一对应。采用 K-均值聚类后，每个簇的组成近似如图 4.5 所示，可以统计出第 i 簇中属于第 i 类的样本，及不属于第 i 类的样本，以及第 i 类错分的样本。其次根据聚类的结果，采用一对多 SVM 方法训练分类器，第 i 个类的分类器训练文本组成为：第 i 类的文本为正类，其他类中所有被错误聚类的文本为负类，从而弥补正负类文本数目的不平衡性。

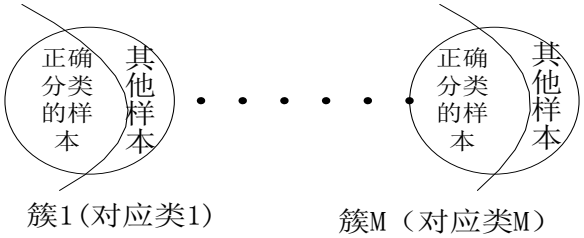

图 4.5 簇样本分布

Fig.4.5 The cluster sample distribution

2. 减小不可分区域

对于不可分区域本文，本文采用一对一 SVM 方法将落在不可分区域的文本进行再次训练，从而减小一对多 SVM 的不可分区域。主要思想是：在上面采用一对多 SVM

方法形成的分类器基础上，将训练集中各类文本作为测试文本通过上面形成的分类器进行分类，将落在不可分区域的文本采用一对一 SVM 进行再次训练，如图 4.6 为具有 3 类文本的训练集训练示意图。

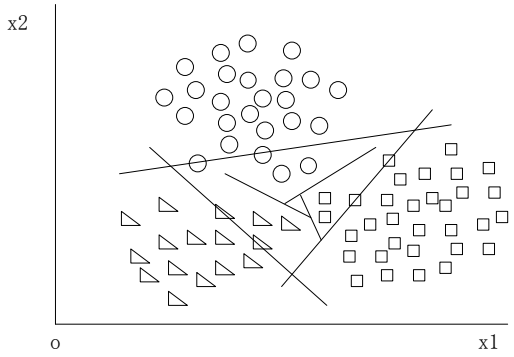

图 4.6 采用一对一 SVM 训练不可分区域样本

Fig.4.6 Train the text of unclassifiable regions by the OVO-SVM

通过以上分析，改进的方法首先采用 K-均值算法对训练文本集做预处理，然后采用一对多 SVM 训练分类器，最后将采用一对多 SVM 方法落在不可分区域的文本采用一对一 SVM 进行再次训练。从而达到解决单独采用一对多 SVM 时存在的训练文本不平衡问题，以及减小了一对多 SVM 方法存在的不可分区域的目的，将这种新方法叫做 Region_Recombine_class 方法。

4.3.2 Region_Recombine_class 算法流程

通过 4.3.1 节的分析，概括 Region_Recombine_class 方法的基本思想为：对有 M 个类的训练集，首先采用 K-均值算法进行聚类，统计各类中不能正确分类的文本。然后进行两步操作，第一步，采用一对多 SVM 方法进行训练，生成 M 个两类分类器，第 i 个分类器的训练文本组成为：第 i 类的文本为正类，其他类中所有不能正确聚类的文本为负类。第二步，将训练集中各类作为测试样本，通过一对多 SVM 产生的分类器进行测试，将落在不可分区域的文本采用一对一 SVM 进行训练。新算法的流程如图 4.7 所示。

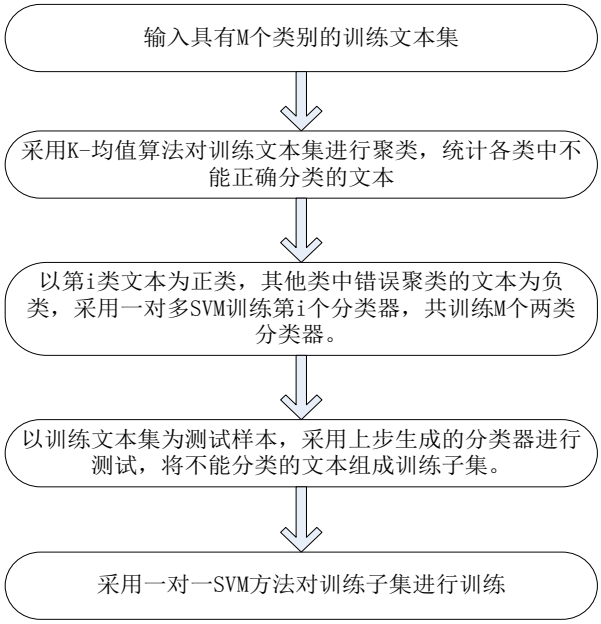

图 4.7 Region_Recombine_class 方法流程

Fig.4.7 The flow chart of Region_Recombine_class

对一个文本进行测试时，首先通过一对多 SVM 方法得到的分类器进行判断，如果此文本是不可分区域的文本，则再通过一对一 SVM 产生的分类器进行判断。

4.4 实验及结果分析

由于本章实验主要在 libsvm 工具箱基础上实现，因此本节将首先对 libsvm 进行讲解，然后对实验结果进行分析。

4.4.1 libsvm 介绍及使用

1. libsvm 介绍

Libsvm 是一个简单、易用、快速有效和开放的 SVM 软件包，是由台湾大学林智仁副教授等设计开发的。该软件包具有强大的功能，他可以直接运行在 windows 环境下，也可以在其他系统上应用。他还提供多种语言的源代码，如 Python、Java、matlab 等，可以方便学习、修改及应用。

该软件包提供了多种 SVM 模型，可以有效的解决多类分类问题和回归问题等，并提供多种核函数以及提供了交互检验的参数选择方法，libsvm 工具包的使用流程如图 4.8 所示。

Libsvm 中的数据格式为：

<label><index1>:<value1><index2>:<value2><index3>…其中<label>是训练数据集的目标值，对于分类它是标示某类的整数，<index> 是以 1 开始的整数，<value>为实

数，也就是常说的自变量。

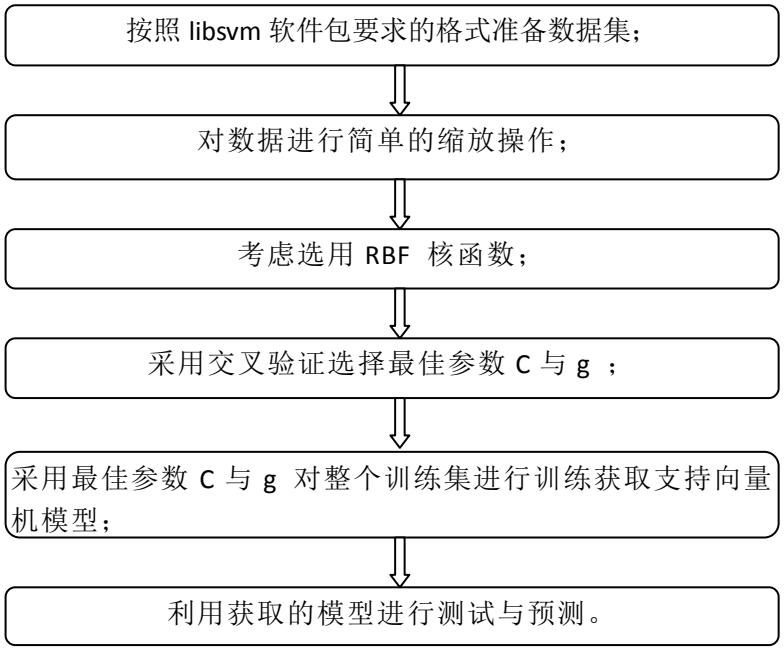

图 4.8 Libsvm 的使用流程

Fig.4.8 The process of use Libsvm

2. 主要函数的使用

（1）svm_train 的使用，该函数主要是用来对训练样本生成分类器的，格式为：svm_train [options] training_set_file [model_file]。其中[options]表示可用的选项，各参数含义如表 4.1 所示。training_set_file 表示训练的数据集；model_file 表示产生的模型文件，模型文件中包含支持向量样本、支持向量样本数、lagrange 系数等必须的参数；该参数可以采用默认文件名，也可以自己设置为常用的文件名。

（2）svm_predict 的使用，svm_predict 是根据得到的模型对测试样本进行预测，格式为：svm_predict test_file model_file output_file.其中 test_file、model_file、output_file 分别表示测试样本、svmtrain 得到的模型文件、输出文件。

（3）svm_scale 的使用，svm_scale 主要是对数据集进行缩放，主要是防止特征值范围大小相差太大，同时在训练连过程中，可以避免因计算核函数而计算内积引起的数值计算困难。

svm_scale 的用法格式为：svmscale [-l lower] [-u upper][-y y_lower y_upper][-s save_filename] [-r restore_filename] filename，其中各参数的含义为：

-l：数据下限标记；lower：缩放后数据下限；

-u：数据上限标记；upper：缩放后数据上限；

-y：是否对目标值同时进行缩放；y _ lower 为下限值，y _ upper 为上限值；

-s save_filename：表示将缩放的规则保存为文件 save_filename；

-r restore_filename：表示将缩放规则文件 restore_filename 载入后按此缩放；filename：待缩放的数据文件（要求满足前面所述的格式）。

表 4.1 svm_train 的参数

Table4.1 the parameters in svm_train

参数名	参数值	意义	默认值
-s:支持向量机的类型	0	c-SVC	0
	1	v-SVC	
	2	单类 SVM	
	3	e-SVR	
	4	v-SVR	
-t:核函数类型	0	线性:u'v	2
	1	多项式	
	2	RBF 函数	
	3	Sigmoid	
-d	——	核函数 degree 设置	3
-g	——	核函数的 gamma 函数设置	1/k
-r	——	核函数中 coef0 设置	0
-c	——	设置 c-SVC、e-SVR、v-SVR 参数	1
-n	——	设置 v-SVC、单类 SVM、v-SVR 参数	0.5
-p	——	设置 e-SVR 中损失函数 p 的值	0.1
-m	——	设置 cache 内存大小，以 MB 为单位	40
-e	——	设置允许的终止判据	0.001
-h	——	是否使用启发式	1
-wi	——	设置第几类参数 c 为 weight?c	1
-v	——	n-fold 交互检验模式，n 为 fold 的个数	2

3. 接口函数介绍

Libsvm 软件包中主要包含：svm.cpp, svm.h, svm_predict.c, svm_scale.c, svm_train.c 等 5 个文件，样本的训练和预测过程主要是通过 svm.cpp, svm-predict.c, svm_train.c 来完成，其中 svm.cpp 中包含了训练和预测时用到的主要类和函数的主要代码，在训练过程中 svm_train.c 调用 svm.cpp 中的 svm_train 函数实现的，svm_train 函数通过接受的参数来处理分类问题，函数调用过程如图 4.9 所示

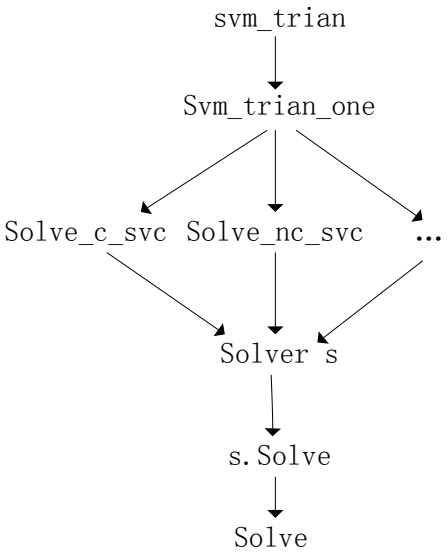

图 4.9 训练过程中函数的调用

Fig.4.9 The function calling process of training

图 4.9 中 Solver s 是调用构造函数，但什么也没做，s.Solver（）函数调用了 SVC_Q 类的构造函数，SVC_Q 类的构造函数中回调 Kernel 类构造函数，并复制目标值（y），同时申请内存，激发 Cache，申请内存，构造双向链表等，最后由 solve 完成剩下的工作。

在 svm.cpp 中还有其他一些常用的函数，例如：

void svm_get_labels（const svm_model*model,int*label）:某类样本的标号；

void svm_destroy_model(svm_model*model):销毁指定的训练模型；

int svm_get_nr_class(const svm_model*model):获取样本的类别；

void svm_predict_values(const svm_model*model,const svm_node*x, double *dec_values):采用训练好的模型预测样本数据目标值；

int svm_save_model(const char*model_file_name,const svm_model *model):将训练好的模型保存在文件中；

void svm_cross_validation:对样本集进行交叉验证；

double svm_predict(const svm_model*model,const svm_node*x):对训练样本进行预测。

4.4.2 实验结果分析

本实验主要在 matlab 环境下，基于 libsvm 工具箱进行，实验数据采用中科院计算所提供的中文文本语料库,该语料库共包括财经、地域、电脑、房产、教育、科技、汽车、人才、体育、卫生、艺术、娱乐 12 大类 14150 篇文本，从中选择财经、电脑、房

产、教育、科技、汽车、艺术、娱乐 8 类作为试验样本，其中训练样本与测试样本数量如表 4.2 所示。

表 4.2 实验样本

Table 4.2 The use of text in experiment

类别	训练样本数	测试样本数
财经	600	100
电脑	800	100
房产	700	100
教育	700	100
科技	700	100
汽车	400	100
艺术	400	100
娱乐	800	100

对实验数据首先进行中文文本预处理，即通过计算权重，特征提取与选择等操作，并整理成 libsvm 能处理的格式，并且一个类的样本是连续存放的。

实验过程中首先采用 k-均值对训练集进行聚类，其中 K=8，并选择每类中的第一个样本为初始聚类中心，将聚类后生成的 8 个簇分别进行统计，并形成新的数据集。K-均值的聚类效果如图 4.10 所示，其中分别计算两个值查全率 $p=c/c_i$ 和查准率 $R=c/v$，其中 i=1,…,k，c 为属于第 i 类且被分到第 i 个簇的样本，c_i 为第 i 类的样本数，m_i 为第 i 簇中的样本总数。

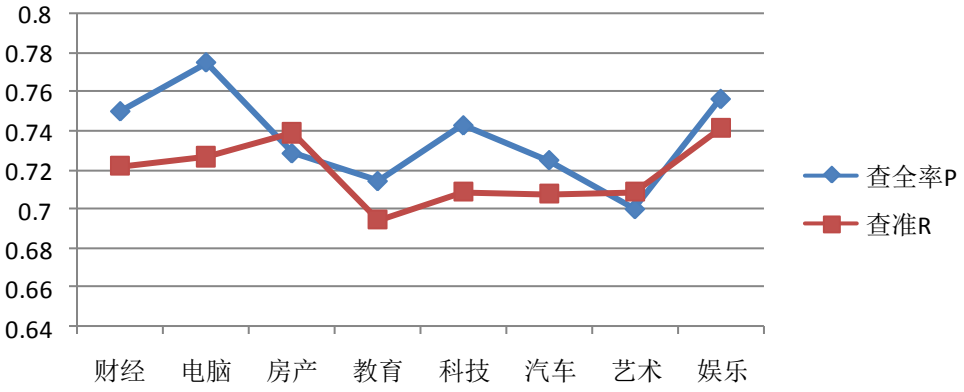

图 4.10K-均值聚类效果

Fig.4.10 The experiment results of K-means cluster

实验分别通过传统的一对多 SVM 方法和本文提出的 Region_Recombine_class

方法进行比较，分类的查准率、查全率、F1 值结果如表 4.3、表 4.5 和图 4.9 所示。

表 4.3 查准率实验结果

Table 4.3 The experiment results of precision rate

类别	一对多 SVM 算法	Region_Recombine_class 算法
财经	0.8333	0.8526
电脑	0.8673	0.8613
房产	0.8118	0.8645
教育	0.8543	0.8738
科技	0.8958	0.9175
汽车	0.8125	0.8673
艺术	0.7767	0.8201
娱乐	0.8113	0.8445

由表 4.3 各类查准率的实验结果，可以看出新方法和传统的 OVRSVM 方法相比总体效果要好，大部分类别只高出两三个百分点，只有少数类别新方法要明显好于一对多 SVM 方法。

表 4.4 查全率实验结果

Table4.4 The experiment results of recall rate

类别	一对多 SVM 算法	Region_Recombine_class 算法
财经	0.8100	0.8300
电脑	0.8500	0.8700
房产	0.8200	0.8300
教育	0.8800	0.9000
科技	0.8600	0.8900
汽车	0.7800	0.8500
艺术	0.7600	0.8400
娱乐	0.8600	0.8700

从表 4.4 各类查全率实验结果可以看出，实验结果与查准率类似，采用新方法时只有少数类别的分类效果明显好于传统的 OVRSVM 方法，大部分类别效果只提高了两三个百分点。

F1 值是反映分类总体效果的评价标准，从图 4.11 可以看出新方法比传统 OVRSVM 算法的分类效果要好，在个别类型下效果明显好于传统的一对多 SVM 方法。

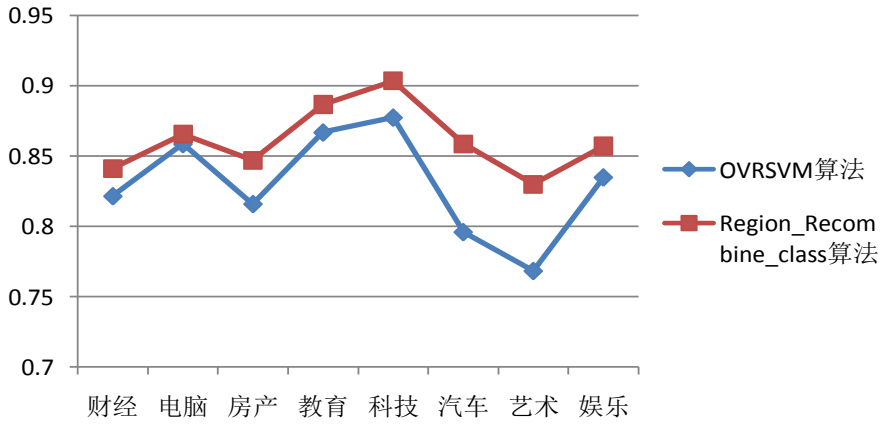

图 4.11 F1 值实验结果

Fig.4.11 The experiment results of F1

从表 4.3、表 4.4 及图 4.11 可以看出 Region_Recombine_class 方法的分类效果总体上要好于传统的 OVRSVM 方法，在个别类别上效果明显好于传统方法，观察效果明显提高的类别，发现这些类别有一个共同的特点，就是类别数目比其他类要少很多，分析这种现象，发现文本数少的类别在采用传统的 OVRSVM 方法时，在训练过程中作为正类的样本要远远小于标为负类的样本，从而导致分类效果差，而采用本文方法可以弥补 OVR 方法这种不平衡现象，而且可以缩小不可分区域。

4.5 本章小结

本章重点研究了 SVM 在多类文本分类问题上的解决方法和基本原理，并详细分析了几种多类分类 SVM 的构造过程，在此基础上结合 K-均值算法和多类 SVM 提出了新的文本分类方法，并通过实验验证了新算法的有效性。

第 5 章 总结和展望

随着网络信息技术的不断发展，各种信息呈爆炸是呈现在人们面前，而这些信息大都以文本形式展现，如何从如此海量的文本信息中快速获取所要的信息是信息时代发展的要求，因此对文本分类的研究越来越重要，也吸引了越来越多的研究人员参与到其中，而在文本分类研究中，对分类算法的研究是其中重要的内容之一，也是保证文本分类的分类速度和效果的关键技术之一，本文就是在这样的背景下对目前常用的文本分类算法进行了研究。

总结

本文主要是在分析了国内外对文本分类研究现状的基础上，结合聚类算法在文本分类中的应用，重点对 K-近邻算法和支持向量机算法进行了研究，本文的主要工作如下：

（1)总体介绍了文本分类的一般过程，并详细讲解了文本分类过程中常用的技术，如：文本的分词，特征提取与特征选择相关方法，常用的聚类与分类算法等等。

（2）对 K-近邻方法在文本分类中的应用进行了深入研究，针对 K-近邻算法存在的类倾斜现象，本文提出结合 SVDD 算法的改进算法。该方法首先在分析 SVDD 的基础上，采用 SVDD 算法对各类进行裁剪，形成新的文本训练集。然后根据新文本训练集的文本分布特点，引入调节因子对 K-近邻方法的判别函数进行改进。该方法有效的解决了 K-近邻算法在分类中的类倾斜问题。

（3)对应用于文本分类的多类 SVM 的各种主要方法进行了研究，如一对一 SVM、一对多 SVM 等。并针对一对多 SVM 在文本分类中存在的样本不平衡性及不可分区域等问题，本文提出了 Region_Recombine_class 方法。该方法首先采用 K-均值算法对训练文本集进行聚类，选用训练集中某类和其他类中错误聚类的文本作为采用一对多 SVM 训练该类分类器的正负文本，从而解决一对多 SVM 训练分类器过程中训练文本数不平衡问题。然后将训练集各类通过一对多 SVM 产生的分类器进行测试，将落在不可分区域的文本，采用一对一 SVM 方法进行再次训练，从而减小一对多 SVM 的不可分区域。

（4）通过一系列的实验证明了以上两种方法应用于文本分类的可行性。

下一步的工作

本文虽然对 K-近邻方法和一对多 SVM 算法进行了改进，但还存在一些问题，并且有许多方面需要进一步的研究，首先采用 SVDD 方法对训练文本进行裁剪时主要考虑的是空间分布，但没有考虑局部文本密度分布的问题，对 SVDD 算法的参数确定也需要进一步的研究，并且 K-近邻算法的判别函数调节因子的计算速度问题也需要进一步的改进。其次在采用将 K-均值聚类算法和 SVM 算法结合生成新的算法时，还需要进一步考虑 SVM 算法的核函数构造方法问题，需要对核函数选择进行进一步的研究。

参考文献

[1] S. Chakrabarti. Hy Pertext databases and data mining. In Proceedings of SIGMOD99,1999.

[2] Bowker R R. Books in Print[M].New York,1997/1998.

[3] 中国互联网络信息中心.第 27 次中国互联网络发展状况统计报告[R].2011.1.

[4] Luhn H P. Auto-encoding of documents for information retrieval systems[M]. Modem Trends in Documentation. New York: Pergamon Press，1959.30-60.

[5] Maron M E, Kuhn J L. On relevance，probabilistic indexing and information retrieval[J].ACM,1960,7(3):216-244.

[6] 李晓黎，刘继敏，史忠植等. 概念推理网及其在文本分类中的应用[J].计算机研究与发展,2000.37(9):1033-1038.

[7] Fuhr N, Hartmana S, Lusting G, Schwantner M, Tzeras K. Air/X-A rule-based multi-stage indexing system for large subject fields[C]. Proceedings of Recherched' Information Assisteepar Ordinateur(RIAO1991).1991:606-623.

[8] Yiming Yang, Jan O Pederson. A Comparative Study on Feature Selection in text Categorization. In:Proc. of the 14th Int'Iconf. On Machine Learning, Nashvile, 1997:412-420.

[9] 肖明，沈英.自动分类研究进展[J].现在图书情报技术.2000,5:25-28.

[10] 李荣陆，王建会，陈晓云等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展. 2005, 42(1): 94-101.

[11] 姚力群,陶卿.局部线性与 One-Class 结合的科技文本分类方法[J].计算机研究与发展. 2005, 42(11): 1862-1869.

[12] 尚文倩,黄厚宽,刘玉玲等. 文本分类中基于基尼指数的特征选择算法研究[J]. 计算机研究与发展. 2006, 43(10): 1688-1694.

[13] 陈晓云,陈祎,王雷,等. 基于分类规则树的频繁模式文本分类[J].软件学报. 2006, 17(5): 1017-1025.

[14] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展[J]. 软件学报. 2006, 17(9): 1848-1859.

[15] 王强,关毅,王晓龙.基于特征类别属性分析的文本分类器分类噪声裁剪方法[J]. 自动化学报. 2007, 33(8): 809-816.

[16] 唐华,曾碧卿.基于遗传算法和信息熵的文本分类规则抽取方法研究[J]. 中山大学学报(自然科学版). 2007, 46(5): 18-21.

[17] 朱靖波,王会珍,张希娟.面向文本分类的混淆类判别技术[J].软件学报.2008, 19(3): 630-639.

[18] 李文波,孙乐,张大鲲.基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报. 2008, 31(4): 620-627.

[19] 郝秀兰,陶晓鹏,徐和祥,胡运发.KNN 文本分类器类偏斜问题的一种处理对策[J].计算机研究与发展. 2009, 46(1): 52-61.

[20] Jain AK,Duin RC. Algorthms for clustering Data[R]. Prentice-Hall Advanced Reference Series. 1988.1-34.

[21] 周宏宇,张政.中文分词技术综述[J].安阳师范学院学报,2010,02:54-56.

[22] 王建会. 中文信息处理中若干关键技术的研究[D].复旦大学，2004.

[23] 沈达阳,孙茂松,黄昌宁.基于统计的汉语分词模型及实现方法[J]. 中文信息, 1998,15(2): 96-98

[24] 刘洋. 中文文本分类中特征选择方法的比较研究[J].计算机与信息技术，2007,03:54.

[25] 周茜,赵明生. 中文文本分类中的特征选择研究[J].中文信息学报，2004,18(3):17-23.

[26] 秦进,陈笑蓉,汪维家等.文本分类中的特征抽取[J].计算机应用，2003,23(2):45-46.

[27]Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communication of ACM, 1975, 18:613-620

[28] 张睿.基于 K-means 的中文文本聚类算法的研究与实现[D].西北大学硕士学术论文.2009:26-34.

[29]李雄飞, 李军. 数据挖掘与知识发现[M]. 北京: 高等教育出版社, 2003:15-35.

[30] 黄文江.中文文本聚类算法分析与研究[D].上海交通大学硕士论文.2010:18-30.

[31] 鲁明羽,李凡.基于权值调整的文本分类改进方法[J].清华大学学报,2003,43(4)：513-515.

[32] 刘彧. 基于贝叶斯理论的文本分类技术的研究与实现[D].吉林大学硕士学术论文，2009:36-46.

[33] 江涛,陈小莉,张玉芳,熊忠阳.基于聚类算法的 KNN 文本分类算法研究[J].计算机工程与应用.2009,45(7):153-158.

[34]石佑红.基于支持向量机的文本分类的研究[D].北京交通大学硕士学位论文,2007:11-24.

[35] 邓乃扬，田英杰.支持向量机——理论、算法与拓展[M]..北京：科学出版社,2009.60-86.

[36] 潘俊辉,王辉.一种基于模糊 VSM 和神经网络的文本分类方法[J].科学技术与工程.2011,11(9):2121-2124.

[37] 蔡崇超.文本分类新方法的研究与应用[D].江南大学硕士论文.2008:5-11.

[38] 何伟成,方景龙.基于信息熵的支持向量数据描述分类[J].计算机应用,2011,31(4):1114-1116.

[39] Pan Zhisong, Ni Guiqiang, Tan Lin,et al. One-classclassification and immune framework in abnormal detection[J]. Journal of Nanjing University of Science andTechnology: Natural Science, 2006, 30 (1): 48-52

[40] 吴定海,张培森,任国全,陈飞.基于支持向量的单类分类方法综述[J].计算机工程,2011,37(5):187-189.

[41] 闫晨.KNN 文本分类研究[D]. 燕山大学硕士学术论文. 2010, p30-37

[42] Chawla N V, Japkowicz N, Kotcz A. Special issueon learning from imbalanced data sets.In: Sigkdd

Explorations Newsletters, 2004, 6(1):126.

[43] Jo Taeho, Japkowicz Nathalie. Class imbalances versus small disjuncts[J]. SIGKDD Explorations Newsletters, 2004, 6(1): 40-49.

[44] Batista E A P A, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. SIGKDD Explorations Newsletters, 2004, 6(1): 20-29.

[45] Guo Hongyu, Viktor Herna L. Learning from imbalanced data sets with boosting and data generation: The Data Boost-IM approach [J]. SIGKDD Explorations Newsletters, 2004, 6(1): 30-39.

[46] Stamatatos E. Author identification: Using text sampling to handle the class imbalance problem [J]. Information Processing & Management, 2008, 44(2): 790-799

[47] Estabrooks A, Japkowicz T H Jo, N. A multiple resampling method for learning from imbalanced data sets [J]. Computational Intelligence, 2004, 28(1):18-36

[48] 孟祥国,多类文本分类的支持向量机网络[D].山东大学硕士论文,2007:12-19.

[49] 孙庆嘉.多类支持向量机的研究与分析[D].北京交通大学硕士学位论文,2010:21-29.

[50] 应伟,王正欧,安金龙.一种基于改进的支持向量机的多类文本分类方法[J].计算机工程.2006,32(16):74-76.

[51] 秦玉平.基于支持向量机的文本分类算法研究[D].大连理工大学博士学位论文.2008:15-34.

[52] 谭冠群.基于多类软间隔支持向量机的文本分类问题研究[D].哈尔滨理工大学硕士学位论文.2008:24-36.

攻读学位期间发表的学术论文

[1] 刘文、吴陈.一种新的中文文本分类算法-one class SVM-KNN 算法[J].计算机技术与发展.已录用.

致 谢

本文的完成首先要特别感谢我的研究生导师吴陈教授，没有他的悉心指导，我不能顺利的完成。他从选题、开题、方案的制定，都给予了我巨大的帮助。研究方法和研究思想的交流给了我很大的启发，对我以后的研究工作出了睿智的分析和指导。他严谨的研究风格和耐心的工作精神也使我受益匪浅。

感谢在论文撰写阶段相互勉励、相互支持、相互帮助的张明华、王万川、孙杰、马言春、桃红、史小五等同学。十分感谢在我完成初稿后史国洁同学在论文排版、论文摘要翻译等各方面给我的帮助。

当然还要感谢我的父母给了我一个良好的成长环境，他们的支持和勉励让我顺利的完成的人生一个非常重要的阶段。再次感谢他们！

最后感谢在我论文研究和撰写过程中给予过我帮助的计算机学院的老师和同学们，在此致以最诚挚的谢意！

基于聚类算法和支持向量机算法的文本分类算法研究

作者： 刘文
学位授予单位： 江苏科技大学