# 第一章 绪论

## 1.1 研究背景与意义

近几年来，随着互联网的蓬勃发展，尤其是移动智能终端的普及，我们的社交、工作、学习、娱乐、出行、购物等都充满了互联网的身影。全世界现在有超过30亿网民，而且这个数量还在不断增加。我国从2013年开始进入4G网络时代，并且规划在10年内进入5G时代，这意味着我们能随时随地使用各种电子设备享受便捷的网络服务。根据2016年6月我国发布的统计报告，我国网民数量已经达到了7.1亿，互联网正在逐渐取代传统媒介称为人们日常生活中获取信息的主要来源。

然而，随着互联网需求的不断增加，网络中的信息量飞速增长。互联网中网页信息80%以上以文本形式存在。文本信息资源作为非常重要的信息类型之一，是人们处理信息资源的重点研究对象之一。想要快速的在互联网中区分和找准信息，就必须对文本信息资源进行分类判断，文本自动分类技术因此得到诞生和发展。文本自动分类是数据挖掘的技术分支之一，作为处理海量信息的重要手段，文本自动分类很大程度上降低查找信息的困难，快速准确地给用户提供想要的信息内容。也就是通过文本自动分类技术，取代传统的、低效的人工分类方法，对文本信息进行快速而准确的分类，提高利用效率。倘若计算机能够帮助人们预处理一些信息的内容判断的工作，那必定会给人们的工作和学习带来非常大的帮助和支持，并将缓解人们当前无法充分利用信息的尴尬情况。

目前，文本自动分类技术已经被其广泛应用，在很多领域都发挥着重要作用，例如搜索引擎、垃圾邮件过滤、主动信息推送服务等。政府可以通过分析公民在社会问题上的在线文本来做出更为公开的决定，公司可以通过分析在线产品评估来识别产品缺陷并预测市场需求，消费者可以通过确定大量在线产品的情绪取向来做出适当的采购决策评论。

因此，针对互联网上海量的、杂乱无章的网页信息资源，利用计算机自动对网页中的文本信息进行分类的必要性就体现了出来，对文本分类技术进行深入研究无论从理论还是实际领域都有重要的意义。

## 1.2 国内外研究现状与发展趋势

文本分类的思想一开始是从信息检索系统演变而来，信息检索是指人们共享不同行业之间的信息时需要提取自己感兴趣的信息。但是，信息的类别十分繁多、冗杂，如何从大量复杂的信息中快速获取有效的信息就变得非常重要。

自动文本分类的研究始于20世纪中期，1958年，美国Luhn提出了采用词频统计来提取摘要的思想，他采用词语的频率与分布信息来估计每个词语的相对重要性。然后再估计每个句子的相对重要度，得分高的句子就被抽取为摘要。1960年，Maron和Kuhns提出了自动关键词分类技术，这是第一篇关于文本分类的文章，对文本分类领域的研究与发展起到了非常深远的影响。他们开创性地提出了朴素贝叶斯分类方法，是一种基于概率的分类方法。在此之后，很多著名的情报学家例如Sparck和Salton等也在文本分类领域取得了卓有成效的研究。

20世纪80年代前，因为技术发展的限制，文本分类大多是基于知识工程的，也就是需要专家人工构建知识工程技术，才能保证分类的有效性。专家制定分类规则是基于知识工程技术分类系统最主要的特点，这些分类规则的重新组合就构成了一个完整的分类系统。由此可以看出，这种分类系统需要大量的专家和知识工程师参与，分类效率和分类准确度都十分低下，而且由于专家研究领域的局限性，这种分类系统的适应性往往也不强。

1990年以来，随着信息技术的快速发展，互联网在人们的生活中占据越来越多的比重，随之而来的信息大爆炸，传统的基于知识工程的分类系统无法处理这种快速增加的数据量。同时，因为机器学习技术的发展，基于机器学习和基于统计的方法正逐渐取代传统方法成为主流的文本分类技术，这也是现在的分类系统的基础。这种分类方法主要通过某种特征选择方法，从训练集中归纳或选择某些特征构成特征集合，以此来创建分类器。这种方法把人们从传统的手工分类中解放了出来，节约了大量的人力物力，成本更低，同时还提高了分类的准确度。

国外的文本分类技术的研究与发展大体可以分为如下三个阶段：第一个阶段进行自动文本分类的可行性与意义的研究，主要时间为1958-1964年；第二个阶段进行实验探索自动文本分类的合理性，主要时间为1965-1974年；第三个阶段进行自动文本分类研究的应用，主要时间为1975-现今；国外的分类系统已经从最初的可行性探索完成到现在的实用化阶段。目前，国外已经有很多可以应于文本分类应用研究的大型商业挖掘软件，例如SAS、SPSS、KXEN等。

我国对于自动文本分类的研究进展要比国外慢很多。一方面，这是因为我国对于文本分类的研究起步较晚。据了解，1981年，南京林业大学的侯汉清教授开启了自动文本分类的探索之路。从基础知识入手，介绍了国外的文本分类技术的发展趋势，清楚地认识到这项技术的重要研究意义。随着技术的发展和研究的深入，中国科学院、清华、北大、哈工大等知名大学都成立了专门研究自动文本分类技术的重点实验室。到目前，我国已经有了基于数据挖掘、大数据分析、机器学习等的文本分类技术。另一方面，因为中文和英文的巨大差异造成的。英文的单词之间有空格分割开来，中文主要是以句分开的，词与词之间则是连续的，没有自然的界限，相比英文分类，中文文本分类要先进行预处理，对连续的词进行分词操作。除此之外，在文本分类中研究英文的语法分析与句法分析的比值要小于中文的语法分析与句法分析之比，这也一定程度上加大了中文文本分类的研究难度。总的来说，我国的文本分类技术的发展也经历了三个阶段：第一个阶段是研究与模仿国外的研究成果；第二个阶段是完善已有的分类技术；第三个阶段是面向中文的发展阶段。

经过几十年的发展，文本分类技术取得了非常大的进步。自动文本分类技术也正在被应用在各个领域，正在影响着人们的生活和学习，有了很好的实践意义和实用的价值，很多优秀的文本分类系统被研究者提出。然而就文本分类而言，仍有可以改进的空间，尤其是特征选择和特征降维方面，还有很多工作可以开展。

## 1.3 本文所做的工作

本文对文本分类技术进行了详细的探讨和研究，包括文本预处理、文本表示模型、特征选择技术、分类器的选择等。由于特征选择技术在文本分类过程中占有着重要的地位，特征选择的好坏直接影响着最终的分类准确度和分类效率，所以本文对特征选择部分进行了详细的介绍，并提出一种混合CHI和MI的改进特征选择方法，克服CHI和MI针对低频词敏感的缺点，提高特征选择的效果。接着本文重点研究了文本分类的算法，介绍主流的文本分类算法，包括朴素贝叶斯算法、K-最近邻算法、决策树算法和支持向量机算法，其中支持向量机算法对海量数据的处理效果最佳，本文也对支持向量机进行重点研究。主要介绍一对一策略下的支持向量机的算法，针对一对一策略算法的只对当前处理的两个类别敏感的缺点，提出使用余弦相似度来计算不同文档之间的相似度，对不同类别文档有所取舍，从而降低类别敏感的影响，最终提高分类的准确性和分类效率。

## 1.4 本文的组织结构

本文主要研究基于机器学习的文本分类技术，所有内容都围绕研究课题展开，具体可分为六个章节，论文的结构以及各章节的主要内容如下：

1. 绪论。主要说明本课题的研究背景与研究意义。接着介绍了本课题的国内

研究现状以及未来的发展趋势，说明国内研究现状及落后原因。最后对本文的组织结构进行了简单的介绍。

第二章 文本分类技术。主要是对文本分类技术的介绍，包括文本分类的基本流程，文本的预处理技术，常用的文本表示模型，以及基本的特征选择技术的介绍，还有常见的分类器选择问题，最后介绍文本分类中方法好坏的评估标准。为后文的研究工作奠定基础。

第三章 混合特征选择方法。主要介绍当前常用的特征选择方法，并分析常见特征选择方法的不足，提出一种改进的混合特征选择方法。其中针对卡方分布特征选择方法和互信息特征选择方法对低频词敏感的缺点，分别引入调节因子，然后混合改进后的两种特征选择方法。

第四章 基于SVM的二分类器处理。介绍SVM中常用的一对一策略，分析一对一策略的缺点，对于不是当前正在处理的两个类别的文档来说，对分类器的影响不是积极的，甚至会影响最终的判断。本文提出了一种使用余弦相似度来判断文档相关程度的方法，去除不相关的文档，提高判断准确率。

第五章 基于机器学习的文本分类系统仿真。通过搭建一个文本分类系统，对前面几章研究的改进方法进行实验验证，分别对比常用特征选择方法和常见的分类器策略与本文研究的方法处理后的分类准确率、召回率和F1值。验证本文研究方法的有效性和实用性。

第六章 总结与展望。针对本文的研究进行了总结和对未来该领域的发展进行了展望。

# 第二章 文本分类技术

## 2.1 文本分类的定义

分类是指将具有相同属性和特征的对象划分为同一个类别的过程和方法。文本分类是指进行分类的对象是文本，预先定义好文本的类别，然后根据待分类文本的内容将文本划分到预先定义好的类别中的过程。文本分类是一种典型的有监督机器学习方法，即分类的类别一开始就已经确定。文本分类的任务可以简单的总结为：在给定了分类体系情况下，根据待分类文本的特征，总结出分类规律。当需要对文本进行分类时，只需要根据总结出的分类规律进行判断，就可以把待分类的文本划分到对应的类别中去。

文本分类的过程类似于数学上的函数映射，分类的过程就是将未分类的文本映射到预定义的类别中的过程。文本分类可以形式化定义为：待分类文档可以用集合D表示，，其中，表示待分类文档 ，n表示待分类文档的个数。预定义的类别用集合C表示，，其中表示类别，m表示类别的种类。利用某种机器学习方法，可以得到一个分类函数f，能够把文档集合D中的每一篇文档都映射到类别集合C中的一个或者多个类别：

 （2.1）

需要注意的是，待分类文档到预定义类别的映射可以是一对一的映射，也可以是一对二的映射。另外，还可以是一对多的映射，一般将多类映射转换为一对二映射进行处理。

## 2.2 文本分类的流程

文本分类一般要经历下面几个过程，分别是文本预处理、特征选择、分类模型训练、分类模型判断，性能评估等。文本分类流程图见图2.1。

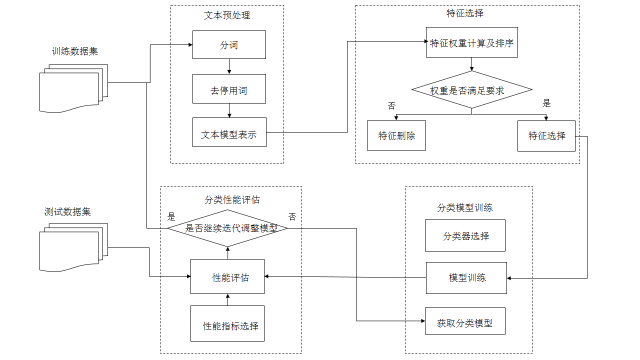


图2.1 文本分类的一般流程

文本预处理部分主要是对文本进行分词，去除停用词等对分类意义不大的信息，比如标点符号、数字和特征字符等，接着使用对应的文本模型进行表示，将文本表示成一种机器可以读的模式，方便后面的处理。特征选择部分主要是先对特征进行降维，降低训练的运算量，然后对特征进行有用性的排序和筛选，构成一个特征子集。分类模型训练阶段主要是对分类器进行选择，使用训练数据集进行模型的训练，得到分类模型，最后使用分类模型对测试数据集进行分类性能评估，根据分类性能评估的结果对分类器进行修正。

## 2.3 文本预处理

文本预处理是文本分类过程中的必要过程，在正式开始进行分类之前，用来将各类文本转换为适当的文本表示形式，主要包括文本标记的处理、分词的处理和去除停用词处理。

### 2.3.1 文本标记的处理

通常来说，文本中除了表示内容的文本信息之外，还包含一些控制文本显示形式的标记、图像、标点符号、声音、动画等媒体信息，还有可能是乱码。这些标记与文本的内容无关，只是对文本的内容进行包装，没有有效的分类信息，对纯文本的分类处理对象来说，这些标记就是噪声信息。在文本预处理阶段，就要去除这些对分类没有贡献的噪声标记。

### 分词处理

在文本分类的过程中，需要先将文本中的句子分割成一个一个单词，然后对单个的单词进行处理，这个过程称为词条化。对英文文本进行词条化处理的时候，主要是根据空格、标点符号就可以达到分割的目的。举个简单的例子，I love text classification.对这句话进行词条化后，变为I / love / text / classification这几个单词。在中文中，由于中文文本的句子中的词与词之间，并没有英文的单词之间有天然的空格分割，所以对中文文本进行词条化需要进行分词。中文文本分词技术一般可以定义为：把连续的中文词组成的句子拆分成一个个含有独立意义的词语。文本分词技术是中英文文本自动分类研究的主要区别，也是信息检索、数据挖掘技术的基础。

中国文字博大精深，相同的词语换个位置，表达的语义就可能发生了变化，句子的形式也是千变万化，所以中文的分词技术也是多种多样。中文分词技术一般是利用某种分词算法，在连续的字或者词组成的句子中添加空格，使之分离成独立的词语。目前中文分词技术主要有三种：基于词频统计的分词技术、基于字典匹配的分词技术和基于理解的分词技术。

（1）基于词频统计的分词技术

在中文中，词语是字的组合，相邻的字在文中出现的次数越多，那么相邻的字就越有可能组成一个词语，所以我们可以用相邻的字与字出现的概率来判断相邻的字是否能够组成一个词语。基于词频统计的分词方法就是以这个理论为依据，将分类字符串中的字的组合频率进行统计，当这个组合频率高于一定的阈值时，就判断字的组合是可以组成一个词语的，否则认为这个字的组合无法组成一个词语。这种分词方法又称为无字典分词方法，因为这种分词方法并不需要字典，只需要统计字的组合的频率即可。也正因为没有字典的存在，这种分词方法只认频率高的组合为词，这就导致经常出现识别出很多频率很高，但并不是有意义的词的情况。除此之外，汉语语言的复杂性和高频使用的词语的数量和类别随着时间会进行变化都导致这种分词方法的时空消耗很大。

（2）基于字典匹配的分词技术

基于字典匹配的分词技术又称为机械式分词技术，该技术把待分类文本中的字符串使用某种方法与字典中的词语进行机械式的匹配，如果在字典中发现了这个字符串，则表示识别出一个词语，如果找不到匹配词语，则将字符串分开作为单个字输出。根据对中文文本字符串的扫描方向来划分，基于字典的分词方法可以分为正向匹配、逆向匹配和双向匹配。根据匹配长度优先策略来划分，基于字典的分词方法可以分为最大匹配和最小匹配。基于字典的分词方法有着分词算法简单、分词效率也较高的优点；但是这种方法的分词正确率主要由字典的完整度和正确性来决定，分词字典的不完备、规则不一致都会导致分词正确率下降，无法处理大规模的文本信息的问题，这是基于字典的分词方法的主要劣势。

（3）基于理解的分词技术

基于理解的分词技术的主要思想是用计算机模拟人的语言理解能力，需要理解句子的含义，对待分类文本进行分词处理的时候，不仅仅只考虑词语是否有含义的情况，还要进行句法和语义的分析，尽量消除歧义的情况。基于理解的分词技术主要包含三个部分：分词子系统、句法语义子系统和总控系统。这种分词方法的一般流程是：利用总控系统协调控制分词操作，句法语义子系统对待分析文本进行句法和语义的分析，最后分词子系统进行分词和歧义情况的处理。基于理解的分词技术需要海量的语言库作为技术支撑，但由于中文语言的复杂性，目前这种方法还处在研究试验阶段，距离实际的应用还任重道远。

### 2.3.3 去除停用词

一般来说，文本的含义主要依靠这段文本中的名词、形容词、动词来表达，而虚词和文本中经常出现但是并不表示文本含义的词就称为停用词。例如，英文中的a、the、and等，中文中的“的”、“了”、“和”等等。这些停用词对表达文本语义没有贡献，而且出现的频率较高，增加了分类算法的处理复杂度。所以为了提高文本分类效率，降低存储空间，我们需要在文本预处理阶段对停用词进行去除。通常对文本进行去除停用词的操作需要使用到停用词表，即将分词后的词语和停用词表中的词进行匹配，如果在停用词表中匹配到了，那么表示这个词为停用词，则要去除，否则，这个词就是有用特征，进行保留。因为去除停用词的正确率和效率都依赖于停用词表，所以停用词表的完备和准确对最终去除停用词的处理结果有比较大的影响。

## 2.4 文本表示及模型

由于文本数据的无结构性，在对其进行分类处理之前需要先将其转化为计算机分类算法便于处理的形式。文本表示主要包含两个方面的含义，一个是如何表示文本的语义特征，二是如何表示文档的语义特征。

### 2.4.1 文本表示

从语法层次来看，一篇文本的特征主要包含词、短语、句子、段落。如何提取这些特征是文本表示的难点。由于句子和段落的语法层次较高，产生的特征维数呈指数增长，一般不采用基于句子和段落的特征。文本表示通常采用词项作为文本的特征项，常见的特征项表示方法有词袋模型（Bag of Words）、短语、N-gram等。

词袋模型是指采用文本中出现的词项作为文本特征项对文本进行表示，词袋由训练集中出现的全部单词构成。词袋模型是文本表示模型中最简单的一种。词袋模型采用的特征都是一个个独立的词语，而文本中的特征除了词语，还有句子、段落和他们之间的语义关系、语法结构等，所以词袋模型的特征项所包含的文本信息相对较少，但模型简单，还是被大部分研究人员采用。

基于短语的文本表示方法是指用文本中出现的短语作为文本的特征项，用短语来表示文档。基于短语的文本表示方法弥补了词袋模型无法处理短语的不足。目前，常用的短语确定方法有两种，一种是句法短语，一种是统计短语。句法短语主要考虑的是词与词之间的语法关联性，统计短语主要考虑不同词在文本中的共线频率。有研究表明，统计短语方法可以提高文本分类的性能。

N-gram方法是指用文本中连续出现的N个词项作为文本的特征项。N的取值一般不超过5，在中文中，N-gram方法一般由相邻的字组成。N-gram方法的优点是不需要进行复杂的预处理操作，不需要海量的词典和复杂的分词程序，具有很强的鲁棒性。在使用同种分类器和特征数目较小的情况下，N-gram方法的分类效果要优于基于词的分类效果。但是，通过N-gram构造特征时，容易导致特征维数过多，增加算法的时间、空间消耗，影响分类效率。

### 文本表示模型

（1）向量空间模型

20世纪60年代，G.Salton教授等人提出向量空间模型（Vector Space Model，简称VSM），因为其形式化表示效果好、应用广泛，现在已经成为最简单的文本表示模型之一。VSM模型首先应用于信息检索领域，随着文本分类技术的发展，VSM在文本分类领域的应用也越来越多。VSM模型用向量的形式表示文本信息，是信息检索领域的经典文本表示模型。

VSM模型的基本思想是从能够表明文本内容主题的拥有独立属性的特征项出发，每个属性构成了特征空间的一个维度，所有的特征项的集合构成一篇完整的文本。用特征向量来表示文本，两个文本之间的相似度通过特征向量之间的相关度来计算。在向量空间模型中，文本集合D由特征项和特征项的权重来表示，形式如：，其中表示特征项，表示特征项对应的权重。两个文本之间的相似度通过对应向量空间模型的向量之间的相关度来计算。常用的文本向量相关度的计算方法有向量内积计算方法和向量夹角计算方法。VSM模型将文本相似度转化为向量之间的相关度，降低了计算文本相似度的复杂度；但是，VSM模型没有考虑文本中特征项的位置等信息，损失了文本结构和语义等信息。

（2）布尔模型

布尔模型（Boolean Model）是一种基于集合论和布尔代数的特征项匹配模型。其主要思想是建立文本对应的特征项集合，根据特征项是否出现在文本中对特征权值进行赋值，如果特征项出现在文本中，该特征项的特征权值就赋值为“1”，否则就赋值为“0”。布尔模型可以说是向量空间模型的一种特例。

布尔模型具有结构简单、快速、易于表达的优点，但是布尔模型使用的是二元判断标准，对于特征项只能判断出现或者不出现，无法对特征项的重要性进行排序，应用效果受到一定的限制。

（3） 概率模型

概率模型是一种基于文本概率的文本表示模型。其主要思想是计算所有文档的概率，然后根据文档的概率对文本进行降序排序，这种情况下的检索效率最高。概率模型利用了特征项与特征项间、特征项与文档间的相关性来进行文本模型表示，解决了向量空间模型和布尔模型的不考虑特征项之间的关系的缺点。

在概率模型中，文本D用特征向量表示，用户查询文本Q用特征向量表示，其中，的权重计算都用二值计算方法，即，，“1”表示特征项出现，“0”表示特征项不出现。文本D和用户查询文本Q之间的概率相关性计算公式如下所示：

 （2.2）

其中，，。F为训练文本集中的文档总数，r为文档集中与用户查询相关的文档数，表示训练集中特征项出现的文档数，表示r个相关文档中包含特征项的文档数。

## 2.5 特征选择技术

文本分类需要大量的文本数据，当对这些文本进行分词操作以后，特征项的维数是非常巨大的。即使经过去除停用词之后，特征维数也经常会达到上万或者几十万维，这么高维的特征向量一方面会加大分类算法学习的时间，降低分类效率，另一方面，很多与文本内容无关的噪声特征的存在也影响了分类准确度。因此，在进行分类算法训练之前，需要对特征向量进行降维处理。特征降维可以去除噪音数据，提高分类的精度和效率。另外，当特征数量过多时，分类器学习过度，会影响分类器在新的数据集上的分类效果，特征降维还可以避免这种过拟合现象。目前，最常用的特征降维技术是特征选择技术。

特征选择方法是指把很多的特征根据某种规则进行排序，选择最能代表文本内容的特征，构成原始特征集合的子集，降低文本向量空间维数，提高分类效率和精度。本文将在第三章对特征选择技术进行详细的阐述。

## 2.6 文本分类技术

前面几节介绍的都是文本分类之前的预处理操作，预处理的目的是形成一个较好的特征空间。然后将预处理后的训练集输入到分类器中进行训练，最后使用测试集进行准确度和效率分析。文本分类技术是文本分类研究中的重点。根据分类规则和类别判断方法的不同，文本分类可以分为两类。

一类是基于规则的分类方法。其基本思想是根据领域专家知识来推倒分类。这种分类方法的基础和前提是需要专家知识，需要专家提前进行大量的规则的定制和修改，才能进行有效的分类操作。基于规则的分类方法有着分类准确率高、分类体系合理等优点。但是，其分类效果和推理规则之间耦合太大，不利于扩展和移植，所以在实际的分类系统中较少采用这种方法。

另一类是基于统计的分类方法。其基本思想是不考虑文本的语义结构信息，假设表示文本的特征词互不相关，通过对训练样本的计算和统计，得到代表文本和类别的关键特征词条。分类器使用选择的特征词条形式化待分类的文本，然后根据训练阶段得到的文本与类别的关系来判断待分类文本的类别。基于统计的分类方法实现简单，通过对训练语料库的分析得到分类知识，分类可信度较高，分类准确度也相对较高。但是，由于这种方法实质上是非确定性的定量推理过程，所以存在对小类别文本忽视的缺点。

文本分类常见的分类器主要有如下几种：

### 朴素贝叶斯分类器

贝叶斯算法是一种以贝叶斯理论为基础的统计学分类方法，是文本分类算法中应用机器学习方法最广泛的算法之一。这里给出贝叶斯定理的公式：

假设、、…、为样本空间的划分元素的集合，且。对任意的随机事件，若，则

 （2.3）

若是把随机事件B看成是结果，那么就是导致结果B产生的种种原因，所以Bayes公式的等号的左边条件概率表示能够产生结果B的原因发生的可能性。Bayes公式等号右边的表示原因独立发生的概率，因为在随机事件抽取之间就可以从已有的数据中得到，所以被称为先验概率。是在历史实验中得到的，表示在事件已经发生的情况下，发生结果B的概率。将Bayes公式应用于文本分类中，即先从训练数据集中收集有用信息，通过训练从而学习到B和的每一种组合的后验概率，得出这些概率以后，按照后验概率的最大值对文本进行分类。

朴素贝叶斯算法（Naïve Bayes），以下简称NB，是贝叶斯算法中最简单有效和应用最广泛的一种。NB算法的特点是有一个独立的假设前提，即在已经给定了目标值的情况下，各个属性变量之间条件相互独立。NB分类器用概率计算，假设待分类文本属于给定的文本集里面的某一类，经过NB分类器可得到

 （2.4）

 （2.5）

然后计算给定的文本集C在输入情况下的概率，得到的最大概率值则表示此类别就是的类别，公式如下：

 if  （2.6）

NB分类算法本质上就是利用先验概率来计算后验概率，即在给出测试文本集以及分类准备的情况下，通过计算和来训练处分类模型。该算法结构简单，在某些场合性能也比较好，但是该算法处理的问题必须符合属性变量之间条件相互独立的假设，而这样的假设在一些实际情况下是不成立的。

### 2.6.2 K--最近邻算法

K—最近邻算法（KNN）是数据挖掘知识范畴内的一种统计算法，它属于消极学习方法，所谓的K—最近邻指的是k个最近邻居，其实就是用某个样本周围距离最近的k个邻居来代表该样本。其原理是对待分类文本不进行训练建模，反而去寻找该测试数据周围距其最近的k个邻居，对这k个邻居的属性值做一个平均计算，并把它当作该测试数据的属性，就此计算出测试文本的类别。

KNN算法的核心就是找到k个近邻的样本节点，而该算法通常是根据以下两个公式来计算待分类的样本与其周围的邻居的距离：

（1）欧式距离，a和b是两个标准化文本向量，计算二者的欧式距离公式为：

 （2.7）

（2）余弦距离，计算标准化文本向量a与b的余弦夹角公式为：

 （2.8）

KNN算法容易理解且实现较为简单，不需要训练建模，有利于应用到数量和规模较小的事件。但是，当文本数据集是不均衡数据集时，该算法缺点也比较明显，输入的测试文本的周围k个邻居可能都是文本量比较大的那个类别，从而产生错误的判断。另外，KNN算法需要对全部待分类的数据集都进行距离的计算，当面对大规模的数据时，计算量太过巨大，计算比率太高。

### 2.6.3 类中心向量方法

类中心向量方法是一种基于向量空间模型的简单分类方法。其主要思想是在训练阶段，通过训练集得到每个类对应的中心向量；在分类阶段，把待分类文本也用向量表示，计算待分类文本向量和每个类的中心向量的相似度，把待分类文本划分到相似度最大的类别中去。如果待分类文本可以属于多个类别，那么可以将相似度进行降序排序，设定一个相似度阈值，把待分类文本划分到相似度大于这个阈值的类别中去。目前，常见的类中心向量方法有：Rocchio方法、Windrow-Hoff方法、EG方法等。

类中心向量方法主要的优点是分类方法简单，分类速度较快。

### 2.6.4 决策树算法

决策树（Decision Tree）是一种有监督学习方法，所谓的有监督学习指的是所有的训练文本集都拥有自己的属性并从属某个类别，然后利用这些训练文本集进行训练得到分类器。二叉树是最典型的一种Decision Tree，二叉树中所有的叶子节点都有且只有一个父节点，每个节点最多可以有两个子节点，每个叶子节点对应着一个类别,每个内部节点都对应一个分割属性，该属性使用分割判断规则来对数据集进行分割，假如是个连续属性，那么的形式是，其中，即是该内部节点n的分割点；假如是一个离散属性，那么的形式是，其中称为该内部节点的分割子集，且，分割标准由内部节点n的分割判断规则和分割属性决定。Decision Tree分类算法一般分成两个阶段，决策树的构建（Building）和决策树修剪（Pruning）。

在构建决策树的过程中，关键在于选择分割方法，不同的分割方法将决策树算法分成两类。第一类是最小Gini指标算法，这种算法常见的应用有CART算法，SLIQ和SPRING；第二类是基于信息论的算法，常见的应用有ID3算法和C4.5算法。常用的Top Down Decision Tree构建算法以节点n、数据集D以及分割方法CL为输入条件，首先对树的节点进行初始化，然后在数据集D中使用分割方法CL来求解节点n的分割标准。如果节点n满足分割条件，则创建节点n的子节点和，选择最好的效果将数据集D分为两个子集和。这样就创建完成，输出以节点n为根节点，节点和为子节点，基于数据集D和分割方法CL的决策树。

目前常用的剪枝方法有三种：悲观剪枝（Pessimistic Pruning）、基于代价复杂度的修剪（Cost-complexity Pruning）和最小描述长度修剪（Minimum Description Length）。1987年，Quinlan提出悲观剪枝，这种方法使用所有的训练集来修剪决策树，但是该方法修剪的决策树过于巨大，且精确度较低。基于代价复杂度的修剪方法使用不同于构建决策树时使用的独立样本集来剪枝，但是在训练集规模比较小的情况下，出于效率的考虑，也可以用来进行决策树的修剪。最小描述长度修剪的基本思路是利用最小长度原理，通过对构建决策树的规则集的精确度进行提升，减小规则集合的消息长度，从而能解决过度拟合的问题。

### 2.6.5 神经网络方法

神经网络（Artificial Neural Networks, NN）分类方法通过模拟人脑神经网络的基本组织特性来完成文本的分类操作。神经网络模型大多为BP网络（Back Propagation Network, BP），即反向传播学习算法。它通常是由三层相同的一系列的神经元构成的层状网络，底层为输入层，中间为隐含层，顶层为输出层，如图2.6.1所示。

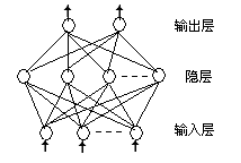


图2-2 神经网络结构图

神经网络分类算法的思路是：在训练阶段，利用训练样本集对神经网络分类系统的每一个输入、输出的连接权重进行调整，直到得到分类效果最佳的神经网络分类器。在分类阶段，分类器针对不同的待分类文本，选择合适的输入层和输出层，从而完成文本分类的判定工作。

目前，常见的神经网络模型有：多层感知机、自适应映射网络等。神经网络分类算法的优点是应用方便、自适应强、获得结果较快较好。但是，神经网络分类方法采用“黑盒”策略，利用了大量的可调参数，分类结果不易解释和理解；另外，分类结果很大程度依赖训练文本集，分类过程很慢，不适合大规模的训练文本语料库的训练学习。

### 2.6.6 支持向量机

支持向量机（Support Vector Machine）是分类效果最好的分类器之一，广泛地应用于分类领域。

## 2.7 分类性能评估

分类评价指标是指在实验中用来评价分类器分类准确度和其他性能的分类指标。在文本分类领域，已经有很多提出的衡量分类器性能的指标，这些指标从不同方面衡量分类器的性能。目前，国际上通用的评价指标有查准率（Recall）、查全率（Precision）、宏平均、微平均等。

### 查准率和查全率

查准率又称准确率，查全率又称召回率，查准率和查全率在检索系统的评价和分类器性能的评价中有广泛的使用。对于文本分类中文本所属类别的情况，使用表2-1计算查准率和查全率：

表2-1

真正属于该类别的文本数 真正不属于该类别的文本数

判断属于该类别文本数 True Positive（TP） False Positive（FP）

判断不属于该类别文本数 False Negative（FN） True Negative（TN）

在表2-1中，TP表示该类别文本被判定为属于该类别的文本数；FN表示该类别文本被判定为不属于该类别的文本数；FP表示不属于该类别的文本被判定为属于该类别的文本数；TN表示不属于该类别的文本被判定为不属于该类别的文本数。

查准率指的是分类器划分给某个类别的文本数量占其真正属于的该类别的本文数的比例，计算公式如xxx：

 （2.9）

查全率指的是真正属于某个类别的文本数量占分类器划分给该类别的文本数量的比例，计算公式如xxx：

 （2.10）

查准率和查全率分别反映了分类器的两个方面的性能，具体采用哪一个评价标准，要根据用户的侧重点来决定。一般来说，这两个指标是互补的，性能较好的分类器应该具备较好的查准率和查全率。

### 值

因为查准率反映的是分类器的准确性，查全率反映的是分类器的完备性，一般情况下，两者相互矛盾，查全率会随着查准率的升高而下降。而F值就是为了综合考虑查准率和查全率而提出的。

值由Van Rijsbergen于1979年提出，值将查准率和查全率综合为一个指标，意图综合考虑分类器的分类性能，两者的相对重要性用表示，的计算公式如下：

 （2.11）

其中，为调节参数，用于调整查准率和查全率在值中的比重。在实际应用中，的取值可以为1，常用的公式如下所示：

 （2.12）

### 宏平均和微平均

查准率和查全率是从类别的角度来评价分类器的分类性能，所以查准率和查全率只代表局部意义。为了评价分类器在整个数据集上的性能，通常对类别评价进行平均，常见的两种平均的方法是宏平均和微平均。

宏平均是每个类别性能指标的算术平均值，强调每个类别对整体结果的影响。宏平均的计算公式如下：

 （2.13）

 （2.14）

微平均是每个文本的性能指标的算术平均值，强调大类别对整体结果的影响。微平均的计算公式如下：

 （2.15）

 （2.16）

### BEP（Break-even point）

由于评价指标查准率和查全率之间时一堆矛盾体，当查准率提升时，查全率就会相对的下降；当查准率下降时，查全率也会相应的提高。为了平衡这两个指标，将查准率和查全率调整到一个相等的值，这个值就是分类器的break-even点。通常采用查准率和查全率的算术平均值来代替BEP的值，计算公式如下：

 （2.17）

BEP的宏平均和微平均计算公式如下：

 （2.18）

 （2.19）

## 本章小结

本章主要介绍了文本分类的概念，介绍了文本分类的流程，接着对文本分类的流程中的各个部分进行了介绍。包括预处理技术、文本表示及其模型、特征选择技术。之后介绍了文本分类中常见的分类算法，包括朴素贝叶斯算法、KNN算法、决策树算法、神经网络算法和支持向量机算法等。最后介绍了分类评估的指标，查准率、查全率、F值和宏平均微平均。

# 第三章 特征选择方法

## 3.1 特征选择方法概述

在文本分类中，需要大量的文本数据来对分类器进行训练，当用向量空间模型对这些分词后的文本进行表示的时候，特征项会非常多，特征维数非常高，后面分类的复杂度很高，分类效率受到影响。实际上，高维特征空间中并不是每一维都对分类有贡献，有很多的模棱两可的特征项不仅对分类没有贡献，相反会增加分类器的学习负担，降低分类效果。例如，“作用”、“过程”、“结果”等等，这种词人工也无法判断是属于什么类别的文本，对分类器来说属于噪声，这种噪声会给分类器算法带来巨大的时间和空间上的计算消耗，也会影响分类器的准确度。相反，像“篮球”、“裁判”、“NBA”等等词语，这类词语很容易看出是属于体育类的文本，这类词语就对分类非常有用。为了降低分类算法计算复杂度，非常有必要对特征空间进行降维处理，目前最常用的特征空间降维技术就是特征选择技术。

特征选择的一般过程是：给定训练文档集合，假设文本分词后得到的所有词语的集合为，用表示集合，特征选择就是从到的一个一对一映射，也就是，接下来以分类器的计算复杂度为出发点，选择一个，选择集合中那些函数值大于或者等于的词语为“选择的特征项”，记为特征集。

目前，常用的特征选择方法有：文档频率（DF）、信息增益（IG）、互信息（MI）、卡方统计量（CHI）等，本章下面将对这些特征选择方法进行详细的介绍。

## 3.2 常见特征选择方法

### 3.2.1 文档频率（DF）

文档频率（Document Frequency，DF）指的是一个特征词条在训练语料库中出现的文档数。文档频率特征选择方法需要先设定一个最小和最大文档频率阈值，对每个特征项的文档频率进行统计，如果特征项的文档频率小于最小文档频率阈值或者大于最大文档频率阈值，那么删除这个特征项，否则保留。文档频率特征选择方法的思想需要有一个假设前提，即如果特征项的文档频率很小，属于低频词，那么就该特征项就没有代表性；如果特征项的文档频率很大，属于高频词，那么该特征就没有区分度。将这些特征项删除对分类效果没有太大影响。

文档频率用DF表示，计算公式如（3.1）所示：

 （3.1）

文档频率选择方法是最简单的特征选择方法之一，因为其和训练语料库的线性复杂度关系，该方法可以应用于大规模训练语料库的统计。但是文档频率特征选择方法的缺点也比较明显，首先该方法认为文档频率过小的特征项不含有或含有很少的类别信息，所以删除这些特征项对分类结果并不会有什么影响。实际上，存在文档频率很低但是却能很好反映文本类别信息的特征项。另外，文档频率特征选择方法只计算特征项是否在文档中出现，而没有考虑特征项在文档中出现的次数，也没有考虑特征项的含义，这是文档频率特征选择方法的主要不足之处。

### 3.2.2 信息增益（IG）

信息增益（Information Gain，IG）概念出自信息论，使用了信息论中熵的概念，通过特征项在文本中出现前后的不同信息熵的差来对特征进行度量。在文本分类领域，信息增益用来度量一个特征的出现与否对分类器判断包含该特征的文本应所属的类别而提供的信息量有多少。

IG特征选择方法的主要过程是：计算每个特征的信息增益，然后按照信息增益的值对特征项进行排序，通过预先设定的特征项的个数或者删除信息增益值小于预定义的信息增益阈值来实现特性选择。对于特征t以及文本所属类别c，IG方法通过对特征t出现在类别c的文本数和特征t不出现在类别c的文本数进行统计，以此来计算特征t对于类别c的信息增益值，计算公式如（3.2）所示：

（3.2）

其中，表示类文档在训练语料库中出现的概率，表示训练语料库中包含特征t的文本的概率，表示文本包含特征t时且属于类别的条件概率，表示训练语料库中不包含特征t的文本的概率，表示文本不包含特征t时且属于类别的概率，m表示类别数。

根据公式（3.2）可以计算出每个特征的信息增益值，并可以按信息增益值对其进行排序，然后可以根据阈值来删除某些特征。总体来说，信息增益方法对文本分类中的特征降维还是有很好的效果，现在已经被广泛地应用于机器学习领域。但是，信息增益方法也存在一些不足，首先信息增益方法需要计算复杂的熵理论公式，计算复杂度高；其次，信息增益方法在计算时考虑了特征不出现的情况，这样会给分类器带来一定的干扰，因为虽然特征的不出现对文本的归属类别判定有一定的贡献，但是特征不出现所带来的负面影响无法被忽略。特别是特征在不同类别的文本中分布不均衡时，那么信息增益方法的效果会大大降低。

### 3.2.3 互信息（MI）

互信息（Mutual Information，MI）的概念也是出自信息论中，用来衡量两个信号之间的关联程度。在文本分类中，互信息用来度量特征词条与类别之间的关联程度。特征词条与类别的互信息计算公式如（3.3）所示：

 （3.3）

其中，表示一篇文本包含特征词条且属于类别的概率，表示包含特征词条的文本数占所有文本数的概率，表示属于类别的文本数占所有文本数的概率。当特征词条与类别没有关联关系时，那么，根据公式（3.3）可知，这种情况下的互信息值为0。当特征词条与类别之间的关联关系很强时，的值远远大于0，当特征词条与类别之间的关联关系很弱时，的值远远小于0。

在实际应用中，一般采用对训练语料库中各类文本出现次数进行统计来模拟互信息的计算。计算公式如（3.4）所示：

 （3.4）

其中，表示包含特征词条且属于类别的文本频数，表示不包含特征词条但是属于类别的文本频数，表示包含特征词条但是不属于类别的文本频数，表示文本语料库中的文本总数。

相比于信息增益方法，互信息方法更加容易理解和实现，思想更加直观。但互信息的缺点在于没有考虑特征词条出现的频度，容易在选择时选择一些低频特征，受特征词条的边缘分布的影响较大。当两个特征的相同时，的大小决定互信息的大小，越大那么互信息值越大，会给分类算法带来负面影响。

### 3.2.4 卡方统计量（CHI）

卡方统计量（Chi-square Statistic，CHI）方法是假设检验的方法，假设特征词条与类别满足一阶自由度的卡方（）分布，然后通过计算特征词条和类别之间的相关程度来进行特征选择。其中，特征词条与类别的相关度与分布的值成正比，统计值越高，说明特征词条含有越多的类别信息，那么该特征词条被选择的几率也就越大。统计量的计算公式如（3.5）所示：

 （3.5）

其中，表示包含特征词条且属于类别的文本数，表示包含特征词条但是不属于类别的文本数，表示不包含特征词条但是属于类别的文本数，表示不包含特征词条且不属于类别的文本数，表示总的文本数，且。

公式（3.5）计算的是单个特征相对于单个类别的统计值，而对于多类分类问题，则有两种常用的方式计算统计值，分别是加权平均和求和。加权平均指的是计算特征词条相对于每个类的统计值，然后取这些值的平均值作为该特征词条的最终统计值。计算公式如（3.6）所示，

 （3.6）

求和指的是计算特征词条相对于每个类的统计值，然后取所有统计值中的最大值作为最终的统计值。计算公式如（3.7）所示，

 （3.7）

## 3.3 改进的混合特征选择方法

### 3.3.1 基于CHI的改进

传统CHI统计方法只考虑了特征词在所有文档集中出现的文档的数量，而没有考虑特征词在某一篇文档中出现的次数，从而夸大了低频词的作用。例如，现在一共有100篇文档，特征词在其中的90篇文档中均出现了一次，特征词在其中的60篇文档中均出现了50次，那么根据CHI公式计算得到，CHI方法会倾向于选择特征词，但实际上特征词比特征词具有更多的分类信息，这是CHI方法的不足，会导致选择的特征不够准确，影响最终分类效果。本文针对CHI方法的这一缺陷引入词频因子进行调节，的计算公式如式（3.8）所示：

 （3.8）

其中，N表示所有的类别总数；表示在类别的文档中，特征词t出现的次数；表示特征词t在所有类别的的全部文档中出现的次数；词频因子表示特征词t在某个特定类别文档中出现的频数占所有类别文档中的总的出现频数的比例。越大，表示该特征在某个特定类别中出现次数较多，而在其他类别中出现次数较少，这样的特征对类别判定能够提供更多的信息。反之，越小，该特征具有越小的类别区分能力，这样的特征词就应该被排除。

### 3.3.2 基于MI的改进

传统的MI方法相比同样是信息论为基础的IG方法，更加的直观且易于理解，计算复杂度也较低。MI方法的一个显著优点是考虑类内不同特征出现的频率，度量的是含有特征t的文本数与类别内总的文本数之间的比率，体现了不同类别内特征对类别的表现能力，有效地利用了文本的类别信息。但是MI方法也存在一些不足的地方，一个主要的缺陷是没有考虑特征本身出现的频度，这会造成MI方法在评估特征时会倾向于选择一些低频特征。从式（3.3）我们可以看出，当两个特征的相同时， 的大小决定互信息的大小，小则互信息大，但实际这样的特征含有的分类信息比较少，从而导致分类效果不理想。本文针对MI的这个不足，提出在原有的公式计算上引入调节参数，改进后的公式如式（3.9）所示：

 （3.9）

其中，通过引入调节参数，添加词频信息，适当增加中高频特征所占比重，降低低频特征的互信息值，避免互信息方法选择过多的低频特征，从而降低低频词对互信息方法的负效用。

不同类别之间，特征的词频也代表了不同的类别区分能力。一个区分能力强的特征词，应该集中分布在某些特定的类别中，也就是不同类别中的特征词频的方差应该尽可能的大，这样的特征含有更多的类别区分信息。为此，本文引入不同类别间特征的词频的方差对MI方法进行优化，记代表特征t在类别j中的频数，则方差可表示为（3.10）：

 （3.10）

其中，m表示总类别的数量。

那么公式（3.9）可以改进为公式（3.11）：

 （3.11）

### 3.3.3 综合度量指标

本文提出混合CHI和MI两种特征选择方法，并对CHI和MI方法各自的不足进行了分析和改进，针对CHI方法不考虑词频的不足，引入词频调节因子。此外引入调节参数减小MI方法对低频词的选择权重，引入类别间特征词频的方差对MI方法进一步优化，使MI方法选择那些集中分布在某些类别的特征，整合两种改进后的方法，得到一种混合的特征选择方法（CHMI）。该方法基于CHI和MI方法，选择那些集中出现在某些类别且在类别文本中分布均匀，出现次数较多的特征。该方法的计算公式如式（3.12）所示：

 （3.12）

## 3.4 本章小结

本章首先对特征选择技术进行了概述，接着介绍了几种常用的特征选择方法，并从原理层面对常见的特征选择方法进行了阐述。根据CHI特征选择方法和MI特征选择方法各自的缺点，对两种方法进行了改进，最终混合两种改进后的特征选择方法，得到综合度量的特征选择指标。

# 第四章 基于SVM的二分类器处理

## 4.1 SVM理论基础

## 4.2 SVM核函数

## 4.3 SVM多分类方法