硕 士 学 位 论 文

基于特征选择和质心构建的文本分类研究

Research of Text Categorization Based on Feature Selection and Centroid Construction

作 者 姓 名：谢华

学科、 专业：计算机应用技术

学　　号：20809349

指 导 教 师：王 健　副教授

完 成 日 期：2010.11

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　　要

随着信息技术的发展，人们能够获取的信息呈现爆炸式的增长。面对日益增多的海量信息，仅仅依靠人工的方式来处理这些信息变得越来越困难。需要一些自动化的辅助工具来帮助人们更好的管理和过滤这些信息。文本分类正是在这样的背景下提出的一种文本自动化处理工具。

文本分类就是将文本集中的每个文本分配到预先定义好的类别集中的某一个类别中去。使用机器学习的方法，其目的就是从实例中进行分类器的学习，然后利用分类器进行自动分类。这是一个有监督的学习问题。当前，存在多种文本分类方法，如朴素贝叶斯，K-近邻，神经网络，基于质心的方法和 SVM 等。文本分类在许多领域，例如网络资源的分类和垃圾邮件过滤等，都得到了广泛的应用。

本文的主要工作是对基于丰富语义信息的文本表示方法进行了研究，并提出了一种新的称为 FSCC 的基于质心的文本分类方法。首先介绍了文本分类的相关背景知识和研究现状。接着详细说明了文本分类的一般流程，包含文本的表示，分类器的选择和训练，最终分类结果的评测。然后研究了文本分类中基于语义信息的文本表示方法。将基于语义的文本表示方法与传统的 BOW 表示方法进行了比较。最后，在传统的基于质心的分类方法的基础上，本文提出了一种改进的基于质心的分类方法 FSCC。在 FSCC 方法中，首先采用特征选择的方法计算特征与类别之间的特征选择值，然后根据特征选择值定义了一个新的质心特征权重计算公式，并由此得到类别的质心向量。最后，采用非归一化的余弦相似度（demoralized cosine measure）来计算文档与质心之间的相似度。本文在不同的语料上进行了实验，实验结果表明，该方法相比经典的质心分类方法以及 SVM，分类效果均有显著的提高。

关键词：文本分类；基于语义；基于质心；特征选择；余弦相似度

Research of Text Categorization Based on Feature Selection and Centroid Construction

Abstract

With the development of information technology, the information people can get are growing in an explosive way. Facing the mass information increasing day after day, people find that dealing with them solely relying on artificial means becomes more and more difficult. People need some automation auxiliary tool to help them management and filter the information more convenient. Text categorization is one kind of text automated tools proposed under such background.

The goal of text categorization is classifying the documents into a fixed number of predefined categories. Using the method of machine learning, its goal is to learn the classifier from examples, and then use the classifier for automatic classification. This is a supervised learning problem. At present, there are many methods for text categorization, such as Naive Bayes, k-nearest neighbor, Neural Network, Centroid-Based Approaches and SVM, etc. Text classification have been widely used in many fields, such as network resources classification and spam filtering, etc.

In this paper, the text representation method based on rich semantic information is studied, and a new method based on centroid-based approach, which is called FSCC, is put forward. Firstly, the background knowledge and research status about text categorization is introduced. Then the general flow of text categorization is given, including text representation, classifier selection and training, the assessments of classification results. And then text representation method based on semantic information in the text classification is studied. The semantic-based text representation methods and traditional BOW representation methods are compared subsequently. Finally, based on the traditional centroid-based classification method, this paper proposes an improved method called FSCC. In FSCC, firstly, the relevancy between features and categories is calculated by using feature selection, and then a new formula for calculating feature weight in a centroid, from which the centroid can be constructed, is defined. Finally, a denormalized cosine measure is employed to calculate the similarity score between a text vector and a centroid. Experiments on different corpus show that FSCC significantly outperforms the traditional centroid-based approach, and state-of-the-art SVM classifier.

Key Words： Text Classification; Semantic-Based; Centroid-Based; Feature Selection; Cosine Similarity

目　　录

1 绪论

1.1 研究背景

随着信息技术的迅速发展，人们获取信息的途径越来越多，能够获取的信息量也越来越大。面对海量数据，如何快速有效的获取人们所需要的数据，成为摆在研究者们面前的一个大问题。搜索引擎的出现大大方便了人们获取信息的方式，只需要简单的输入关键字，人们就能找到自己需要的信息。但是，搜索引擎返回的结果与人们的期望之间还存在一定的差距。一方面，搜索引擎的结果往往不能令人满意，存在大量无关的搜索结果。另一方面，随着信息量的不断增加，搜索引擎的速度也成为了影响其性能的瓶颈。人们希望能将信息分门别类地进行组织，这样一来，一方面可以根据用户的兴趣直接到对应的类别中查找，因此，可以提高搜索的精度。另一方面，确定对应的类别之后，就能大大缩小搜索的范围，因此，也能提高搜索的速度。这正是文本分类、聚类的研究动机之一。

文本分类就是将文本集中的每个文本分配到预先定义好的类别集中的某一个类别中去。最初，文本分类是通过人工的方式进行的。由相关领域的专家对文本的类别进行判定。这样做不仅效率低下，而且随着文本数据的不断增长，采用人工的方式进行文本分类变得越来越不可行。

直到 20 世纪 50 年代末，自动文本分类研究才真正开始，这一研究领域的先驱是 H.P.Luhn，他首先提出词项信息统计的思想。20 世纪 60 年代初，Maron 发表了有关自动文本分类的第一篇论文[1]，随后，这一领域的研究越来越多，许多著名的研究者，如 Sparck 和 Salton 等都进行了比较深入的研究[2]。到 20 世纪 80 年代末，出现了采用知识工程的方法来建立专家系统，并以此来建立自动文本分类系统的方法。即在进行自动文本分类的时候，引入专家规则。这种方法的优点是结果容易理解，但费时费力，且结果的准确性通常比较低。更重要的是，随着规则集的不断增加，难以保证规则集的一致性。进入 90 年代以后，随着信息检索相关技术的不断发展，基于机器学习和统计自然语言处理的文本分类受到越来越多人的重视。基于机器学习的方法的准确性一般比较高，且来源于真实的文本，可信度也比较高。其中，朴素贝叶斯[3]、K-近邻[4]、神经网络[5]、基于质心的方法[6-9]和支持向量机[10]等机器学习算法得到了广泛的使用。基于机器学习的文本分类方法能利用计算机帮助人们自动的进行文本分类。因此，能大大提高文本分类的效率，同时也使得海量文本数据的分类问题成为了可能。

当前，文本分类在很多领域，如网络资源的分类[11]，Blog 分类[12]以及垃圾邮件的过滤等领域，都得到了广泛的应用。

1.2 研究现状

文本分类包含三个阶段，即文本的表示、分类器的选择与训练、分类结果的评测。其中，文本的表示以及分类器的选择都对最后的分类效果有很大的影响。因此，当前的文本分类方法都试图在这两个阶段进行改进，以提高分类的效果。

在文本的表示阶段，需要将文本转化为计算机能处理的形式。经典的文本表示方法是向量空间模型（VSM）。在向量空间模型中，文档被表示成向量的形式。向量空间模型建立在特征独立假设基础之上，即文档中的特征词与特征词之间相互独立。因此，它忽略了文本中存在的语义信息。针对这一问题，提出了许多向文本表示中加入语义信息的方法[13-15]。例如，文献[14]通过引入维基百科（Wikipedia）作为外部资源来构建得到一个语义矩阵，对每篇文档，从语义矩阵中挖掘出语义信息，并将该语义信息加入到文本的表示中，以达到丰富文本表示的作用。文献[15]通过发掘文本中的隐含的multi-grams，并将这些 multi-grams 作为特征加入到文本的表示中，从而达到丰富文本表示的目的。相比原始的 BOW（bag of words）模型，由于这些方法引入了一部分语义信息，因此，在一定程序上提高了分类的精度，但往往代价比较大。

在文本的分类阶段，选择不同的分类方法，对最终分类结果也会产生影响。目前，存在多种基于机器学习的文本分类方法，如朴素贝叶斯、K-近邻，支持向量机、基于质心的方法等。其中，支持向量机（SVM）作为一种分类方法，由于它能有效的处理文本分类中面临的一些普遍问题，即高维的特征空间、文档向量的极度稀疏等[16]。因此，SVM 是各种分类方法中分类效果最好的方法之一。

基于质心的方法作为应用最普遍的一种分类方法之一，具有简单的特点。但由于它是基于文档应当被划分到与该文档最相似的类别中这一假设之上，而这一假设往往不成立，从而导致对模型的敏感性，因此，其分类效果往往远低于其它的分类方法，如 SVM。针对基于质心方法效果较差这一问题，提出了许多改进方法[6-9]，它们大多是利用词项在文本集中的分布信息来构建类别的质心。其中，文献[6]提出了一种称为 CFC 的基于质心的方法，它结合了词项在类别内部和类别之间的分布信息来构建质心向量，在 Reuters-21578 和 20-newsgroup 两个语料上都取了很好的效果，其分类的效果要好于 SVM 分类方法。

1.3 本文的工作

由于文本分类系统中文本的表示，分类器的选择都会对分类结果产生影响，因此，本文主要对这两个方面进行一些研究。本文的主要工作有两方面，一方面是调研了当前向文本中加入语义信息的两种方法，即隐式加入语义信息和显式加入语义信息。前者的典型代表是潜在语义索引（LSI），而后者通常是采用维基百科作为外部资源向文本表示中加入语义信息。针对前人的工作，设计了一个简单的向文本表示中加入语义信息的方法，并进行了相关实验，将基于语义的分类方法与传统的 BOW 表示方法进行了比较。

另一方面，在分类器的选择方面，本文重点研究了基于质心的分类算法。在传统的基于质心的分类方法的基础上，提出了一种改进的基于质心的分类方法 FSCC。在 FSCC 方法中，首先采用特征选择的方法计算特征与类别之间的特征选择值，然后根据特征选择值定义了一个新的质心特征权重计算公式，并由此得到类别的质心向量。最后，采用非归一化的余弦相似度（denormalized cosine measure）来计算文档与质心之间的相似度。本文在不同的语料上进行了实验，实验结果表明，该方法相比经典的质心分类方法 AAC 和 CGC，改进的质心分类方法 CFC 以及经典的分类方法 SVM 等，分类效果均有不同程度的提高，对某些方法的效果有显著的提高。最后，还设计了其它的一些构建类别质心的计算方法，并与 FSCC 方法进行了比较。

1.4 本文的结构

论文共分为四章，详细介绍了文本分类的基本流程，基于语义和基于质心的文本分类方法，实验设计、结果和分析等。具体章节安排如下：

第一章，绪论，综述了论文研究工作的背景以及研究现状，介绍了本文研究的主要研究工作和论文的结构安排。

第二章，介绍了文本分类的主要流程以及实验语料，实验工具和评测指标。

第三章，详细介绍基于语义的文本分类方法，以及实验设计与结果分析。

第四章，详细介绍基于质心的文本分类方法，以及实验设计与结果分析。

论文的总结部分，介绍了本文的研究内容，主要工作及下一步的工作。

2 文本分类的流程

2.1 文本的表示

2.1.1 文本预处理

通常情况下，待分类文本并不是标准的结构化数据。有些语料根本没有相关的标签用来标识正文、摘要等信息，有些语料虽然有标签，但标签并没有严格按照特定的格式（如 XML）来组织。这些都会给后续的文本处理带来不便。因此，在文本表示阶段，首先需要对语料集中的文本进行预处理，将文本转化为统一的格式[17]，如 XML 格式，以便于后续的统一处理。

文本的预处理还包括对文本的内容，即文本中的特征词，进行一些处理。对英文文档，需要进行词干化、去停用词等处理。对中文文档，则需要进行分词、去停用词等处理。所谓去停用词，就是去掉那些在文本中经常出现、对于分类没有任何作用的词。如“的”、“有”、“我们”等。这些词没有任何信息量，应当预先被过滤掉。通常的方法是先建立一个停用词表，对于文档中的词，如果属于停用词表中的词，就过滤掉。这样就能去除不含信息量的噪音特征，而只保留了那些对于分类有作用的词。

对于英语文档，由于一个词有很多种变体。因此，对于文档中的每个词，应该进行词干化处理，即将词的所有变体都转化为词根的形式。如：“studies”、“studying”和“studied”都是单词“study”的变体，在遇到这些词时，应该统一转化成词根“study”的形式。这样做，不仅有助于分类结果的提高，还能起到空间降维的作用。当前，存在不少词干化的算法，利用这些算法，可以直接帮助我们进行词干化处理。其中，使用比较广泛的算法如 Porter Stemmer。

对于中文文档，由于中文文本词与词之间没有明显的界限，因此，需要进行分词。将文档中的句子切分成一个个独立的词语。中文分词方法有很多种，主要分为基于字符串匹配的分词方法，基于理解的分词方法和基于统计的分词方法这几种。其中，基于字符串匹配的分词方法又包括正向最大匹配法、逆向最大匹配法和双向最大匹配法。在这些分词方法的基础上，已经有一些成熟的中文分词工具，很多也都取得了不错的结果。如：哈尔滨工业大学开发的 IRLAS 词法分析系统。这一系统提供包括分词，词性标注，以及基本的命名实体识别功能[18]。

2.1.2 向量空间模型

经过文本的预处理阶段，文本被转化为词序列。由于要使用机器学习的方法进行文本的自动分类。因此，首先必须将文本转化为机器能够识别和处理的形式。这涉及到文本分类中的文本表示问题。目前，使用最多的文本表示方法是向量空间模型（Vector Space Model，VSM）。

向量空间模型的基本思想是把文档 doc_i 看作是向量空间中的一个 n 维向量（w_{i1}，w_{i1}，w_{i2}，w_{i3}，…，w_{in}）。其中，w_{ik}，k=1，2，…，n 表示特征词 w_{ik} 在文档 doc_i 中的权重。在向量空间模型中，特征权重的计算是关键。当前，存在多种特征权重计算方法。经典的计算特征词权重的公式是 tfidf。计算公式是：

$$tfidf_{t,d} = tf_{t,d} * \log\left(\frac{N}{n_t}\right) \qquad (2.1)$$

其中，$tf_{t,d}$ 表示词条 t 在文档 d 中的出现的次数；n_t 表示词条 t 在文档集中出现的文档数；N 表示文档集中包含的文档个数。

词项权重通常是词频的函数。一般认为一个词项在一篇文档中出现的次数越多，说明该词项对于这篇文档具有比较大的区分性。因而，应当被赋予较大的权重。在式 2.1 中，$tf_{t,d}$ 正是词项在文档中的出现次数对于该文档区分度的体现。另一方面，如果一个词在所有的文档中都出现过，则说明该词对于文档的区分性很低。因而，应当被赋予较小的权重。式 2.1 中对 n_t 取倒数，正是基于对这一观点的考虑。

向量空间模型得到了广泛的应用，它大大简化了文档中特征与特征之间的关系，大量的实验结果也表明该方法往往能取得很好的效果。但是，向量空间模型是基于特征独立假设基础之上的，它忽略了文档中特征词与特征词之间的语义联系。然而，实际上，文档中词与词之间，特别是相邻的词与词之间，通常存在一定的语义关系。因此，向量空间模型必然会导致语义信息的丢失。这也是向量空间模型存在的不足。

2.2 特征降维

2.2.1 特征选择

文本分类中存在一个主要问题就是高维的特征空间[16]。特征空间的维数经常达到上万维。高维的特征空间一方面会引入很多噪音特征，另一方面又会导致较高的计算代价。因此，在文本处理过程中，经常需要进行特征降维（aggressive dimensionality reduction）处理。

特征降维的主要方法有两种，一种是特征选择，另一种就是特征重构。

所谓特征选择（Feature Selection），就是从一组特征中挑选出一些“最有效”的特征以达到降低特征空间维数的目的。这里的“最有效”，是指特征具体很好的分类信息。由于衡量特征有效性的方式多种多样，从而产生了多种特征选择方法。主要的特征选择方法包括文档频率（Document Frequency, DF）、信息增益（Information Gain, IG）、互信息（Mutual Information, MI）和开方检验（X^2-test，CHI）[19-22]等。

（1）文档频率（DF）

词项的文档频率是指特征在文档集中出现的文档数，它是最简单的一种特征选择方法。采用文档频率作为特征选择方法是基于以下观点：文档频率低于某个阈值的词属于低频词，它们不含或含有较少量的类别信息[19]；而文档频率高于某个阈值的词属于高频词，也没有类别区分度。将这些词过滤掉，不仅能达到空间降维的目的，还有可能提高分类结果的精度。

文档频率的计算公式是：

$$DF(t)=\frac{n_t}{N} \tag{2.2}$$

其中，n_t 表示词条 t 在文档集中出现的文档数；N 表示文档集中包含的文档个数。

实际中，并不单独使用文档频率，而是将文档频率方法与其它特征选择方法结合起来使用。例如：先使用文档频率过滤掉一部分特征，然后再进一步使用其它的特征选择方法进行特征选取。

（2）信息增益（IG）

信息增益是特征选择方法中使用较多的一种方法。它从信息论的角度出发，以各特征取值情况来划分学习样本空间，根据所获信息增益的多少来筛选有效特征[20]。在进行信息选择时，选择信息增益大的那些特征。

信息增益用到了信息论中有关信息量（就是“熵”）的概念，特征 t 的信息增益值为不考虑任何特征的熵与考虑该特征后的熵的差值。信息增益的公式如下：

$$\begin{aligned} IG(t)&=H(C)-H(C\mid T)=-\sum_{I=1}^{n}P(C_i)\log_2 P(C_i)+\\ &P(t)\sum_{i=1}^{n}P(C_i\mid t)\log_2 P(C_i\mid t)+P(\bar{t})\sum_{i=1}^{n}P(C_i\mid \bar{t})\log_2 P(C_i\mid \bar{t}) \end{aligned} \tag{2.3}$$

其中，P（C_i）表示类别 C_i 出现的概率；P（t）表示特征 t 出现的概率；P（$C_i\mid t$）表示出现特征 t 时类别 C_i 出现的概率。

信息增益被公认为是一种较好的特征选择方法[21,22]。

（3）互信息（MI）

互信息是信息论中的概念。在特征选择领域，特征 t 和类别 c 的互信息体现了特征与类别之间的相关度。某个类别中出现的概率高，而在其它类别中出现的概率低的特征将获得较高的互信息。

特征 t 和类别 c 的互信息定义如下：

$$MI(t)=\log_2 p(t\mid c)-\log_2 p(t)=\log_2 \frac{p(t\mid c)}{p(t)} \tag{2.4}$$

其中，p（$t|c$）表示在类别 c 中特征 t 出现的概率；p（t）表示特征 t 出现的概率。

（4）X^2 统计（CHI）

X^2 统计量可以用来衡量特征 t 类别 c 之间的统计相关性强度。我们感兴趣的是那些与各个类有强相关联的项。

特征 t 类别 c 之间的 X^2 统计值 $X^2(t,c)$的计算公式如下：

$$\chi^2(t,c)=\frac{N\times(A\times D-B\times C)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{2.5}$$

其中，A 是属于类别 c 并且包含特征 t 的训练文档个数；B 是不属于类别 c 且包含特征 t 的训练文档个数；C 是属于类别 c 且不包含特征 t 的训练文档个数；D 是不属于类别 c 且不包含特征 t 的训练文档个数；N 是训练文档总数。

2.2.2 特征重构

另一种特征降维的方法就是特征重构。与特征选择方法不同，特征重构并不是简单的从原来的特征集合中选择一部分特征。而是将原来的高维特征空间映射到一个新的低维特征空间。新的特征空间的某一维可能是原来特征空间某几维的线性组合。这种方法不但可以起到特征降维的作用，还能挖掘词项与词项之间内在的语义关系，从而得到低维和反映词项关系的特征。特征重构的代表方法是潜在语义索引（Latent Semantic Index，LSI）。

潜在语义索引是一种根据词条的共现信息探查词条之间内在的语义联系的方法。通过对文档矩阵进行特殊的矩阵分解，将矩阵近似地映射到一个低维的潜在语义空间上。潜在语义索引是矩阵的 SVD 分解在信息检索领域的应用。

潜在语义索引把同现的词条映射到同一维空间上，而非同现的词条映射到不同的空间上。潜在语义空间与原来的空间（VSM）相比，空间维数要小的多。因此，能起到特征降维的作用。

2.3 分类器的选择与训练

2.3.1 基于机器学习的分类方法

基于机器学习的方法现已成为文本分类领域主流的分类方法。机器学习其实是让机器去模仿人的学习过程。当人们对一件事情作出判断时，往往是依据自己已经掌握的知识对事情进行分析和总结，然后作出自己的判断。同样，要让机器处理一个任务，必须先让机器“学习”一些知识，机器基于已经学习到的这些知识，对要处理的任务作出自己的“判断”。

对基于机器学习方法的文本分类问题，由于要提供训练集样本，因此，文本分类属于有监督的机器学习问题，也属于机器学习中的模式识别领域。当前，已经有不少机器学习算法在文本分类领域得到了广泛的应用，并取得了很好的分类效果。以下几小节介绍在文本分类中使用比较广泛的几种机器学习算法，即 SVM 算法，朴素贝叶斯算法和 K-近邻算法。

2.3.2 支持向量机（SVM）

Vapnik 于 1995 年发明了支持向量机（Support Vector Machines, SVM）这一机器学习算法[23]，支持向量机的提出是统计学习理论领域的重大突破，它主要用来解决二分类模式识别问题。支持向量机方法建立在统计学习理论的 VC 维理论和结构风险最小化原理基础之上[18]，在样本信息有限的条件下，在模型的复杂性和学习能力之间寻求最佳的折衷，目的是得到最好的泛化能力。Joachims 最早提出将支持向量机方法应用于文本分类领域[16]，并验证了这一方法能取得很好的效果。

从几何学的观点来看，SVM 的目的就是要在 n 维空间中寻找最优的分类超平面（决策超平面），将正例和负例分开。所谓最优分类超平面，是指该超平面在正确划分正例和负例的前提下，和它们的距离（即分类间隔）要尽可能的大。

下面从两类线性可分的情况开始，然后再扩充到一般的不可分数据的情况。

对两类线性可分的情况，图 2.1 给出了一个示例。SVM 的基本思想可以采用图 2.1 来说明。

图 2.1 中两种类型（实心和空心）的样本点分别代表正例和负例样本。超平面 H 即为最优分类超平面，超平面 H1 和 H2 上的点为支持向量。可以看到，超平面 H 是同时离正例和负例的支持向量的距离最远的平面。

设 x_i，i=1,2,…,N 是训练集 X 中的特征向量。这些向量来自于类 w_1，w_2，并且假设是线性可分的。目的是设计一个超平面，将所有训练向量正确分类：

$$g(x)=\omega^T x+\omega_0=0 \qquad (2.6)$$

对测试集实例 x 的分类是依据 $w^{\mathrm{T}}x+w_0$ 的值来决定的。分类的规则可以表示为：

$$g(x)=\omega^T x+\omega_0\geq 1 \quad y_i=+1 \qquad (2.7)$$

$$g(x)=\omega^T x+\omega_0\leq 1 \quad y_i=-1 \qquad (2.8)$$

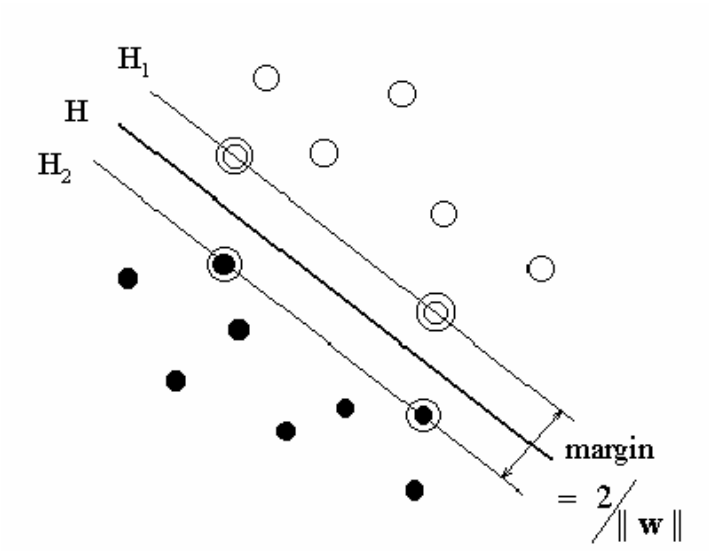

图 2.1　SVM 方法中的线性可分问题

Fig. 2.1　Linear separable problem for SVM

对于非线性问题，通常的解决方法是定义一个核函数，通过核函数（非线性变换）实现空间的映射，即将线性不可分离问题转化为某个高维空间中的线性可分问题，在变换得到的空间中寻找最优分类超平面。

Joachims 指出[16]，SVM 方法在以下两个方面要优于其它的文本分类方法：

（1）对 SVM 方法，特征选择已无必要。这与其它的分类方法不同。因为特征选择的主要目的是为了降维（reduc）和解决过拟合（over fitting）问题。而 SVM 能有效的克服特征空间的高维问题，当出现过拟合时也能表现出很好的鲁棒性；

（2）不需要在验证集（validation set）上通过工人或机器的方法来调节参数。SVM 方法会自动的选择参数，以取得最好的结果。

SVM 方法具有坚实的理论基础，并且它能有效的解决向量空间的高维和特征的极度稀疏，因此，SVM 是目前的机器学习方法中效果最好的算法之一。

2.3.3 朴素贝叶斯分类方法（Naive Bayes）

朴素贝叶斯（Naïve Bayes，NB）分类算法是一种基于概率统计的机器学习算法[18]。NB 通过统计词项和类别之间的联合概率来估计给定文档属于类别的概率。NB 方法的朴素（Naïve）部分在于它的词项独立性假设，即不同词项在给定类别下的条件概率分布是互相独立的。这一假设使得 NB 分类器不需要计算词项之间的联合分布概率，大大简化了概率分布的计算，因此，其速度远快于非朴素贝叶斯的方法，后者为指数复杂度。

假设 $C=\{c_i| i=1,2,\ldots,m\}$为给定的类别体系，分类任务要求，对于给定的文档 d，判断其所属的正确类别。完成这个任务只需要分别计算文档 d 在每个类别中出现的后验概率 $P(c_i|d)$，然后将之分类到概率最大的类别中去即可，形式化表示如下：

$$c^* = \arg\max_{c_i \in c} p(c_i \mid d) = \arg\max_{c_i \in c} \frac{p(c_i)p(d \mid c_i)}{p(d)} \tag{2.9}$$

$$\propto \arg\max_{c_i \in c} p(c_i)p(d \mid c_i) \tag{2.10}$$

$$= \arg\max_{c_i \in c} p(c_i)\prod_{j=1}^{n} p(w_j \mid c_i) \tag{2.11}$$

其中式 2.10 到式 2.11 转化过程中，做了如下的独立性假设：文档中所有词之间相互独立，也即将文档看作是“bag of words”，正是由于做了这个很 naive 假设，因而才称这种方法为 Naïve Bayes, 其中 Bayes 体现在式 2.10 中 Bayes 公式的采用。

到此，任务转化为如何合理的计算 $P(w_j|c_i)$，即每个词在各个类别中出现的后验概率。而 $P(w_j|c_i)$的计算可以采用很多方法，在给定训练集的情况下，最简单的处理方法是根据词 w_j 在各个类别中出现的频率来近似地估计，如下所示：

$$p(w_i \mid c_j) = \frac{n_{ij}}{n_j} \tag{2.12}$$

其中，n_{ij} 表示在类别 c_j 中出现特征 w_i 的文档数，n_j 表示类别 c_j 包含的文档数。

式 2.12 可能发生零概率问题，这可以通过采用简单的平滑方法解决，如加一平滑等。采用加一平滑后的计算公式如下（n 为总的特征数目）：

$$p(w_i \mid c_j) = \frac{n_{ij}+1}{n_j+n} \tag{2.13}$$

朴素贝叶斯的特征独立性假设虽然一般情况下并不成立，但它具有很好的鲁棒性，在处理真实数据时往往具有很好的效果[24]。