# KNOWLEDGE DISCOVERY

# 知识发现

## 科技文献内容挖掘技术研究

吉久明　李　楠　著



上海科学技术文献出版社
Shanghai Scientific and Technological Literature Press

KNOWLEDGE DISCOVERY

# 知识发现

## 科技文献内容挖掘技术研究

吉久明　李　楠　著

上海科学技术文献出版社
Shanghai Scientific and Technological Literature Press

**图书在版编目（CIP）数据**

知识发现：科技文献内容挖掘技术研究 / 吉久明等著．—上海：上海科学技术文献出版社，2017
ISBN 978-7-5439-7433-3

Ⅰ．① 知… Ⅱ．①吉… Ⅲ．①科技文献—知识管理 Ⅳ．① G257.36

中国版本图书馆 CIP 数据核字 (2017) 第 126231 号

责任编辑：应丽春
封面设计：袁　力

**知识发现：科技文献内容挖掘技术研究**
吉久明　李　楠　著
出版发行：上海科学技术文献出版社
地　　址：上海市长乐路 746 号
邮政编码：200040
经　　销：全国新华书店
印　　刷：常熟市文化印刷有限公司
开　　本：787×1092　1/16
印　　张：9.75
字　　数：225 000
版　　次：2017 年 7 月第 1 版　2017 年 7 月第 1 次印刷
书　　号：ISBN 978-7-5439-7433-3
定　　价：48.00 元
http://www.sstlp.com

# 内容简介

*Introduction*

本书主要研究基于学术文献的知识发现问题。基于非相关学术文献的知识发现问题包括：分类、特征提取、术语或实体提取、术语或实体关系提取及本体构建等，其中术语或实体关系提取一般作为领域本体构建的基础工作。由于领域本体收集了领域内相关概念及概念间关系，为基于文献的语义推理提供了可能，因此，领域本体构建工作又为基于文献的知识发现提供了广阔的发展空间。本书提出了一种基于支持向量机的学术论文影响力预测方法，基于单分类器的科技论文多分类问题解决方案，设计了一种科技论文的特征提取方法，进而提出了提取领域框架语义网络本体核心词汇、语义配价模板及例句选择方法，并进行了实验验证，且以导电塑料制备领域大量科技文献为应用背景，对该领域的框架语义网络本体构建过程进行分析，以验证上述面向科技文献内容的知识发现方法的有效性和正确性。

本书可供计算机应用相关专业本科高年级学生、硕士研究生、博士研究生教学和科研使用，还可作为计算机应用专业各专业研究生和本科生的选读教材，同时也可作为从事相关研究和实际工作的高等院校教师、科研工作者的参考书。

# 前言

*Preface*

知识是一种特殊信息，是人对信息加工、吸收、提取、评价的结果。这种特殊信息与本能有着本质的区别，随着时间推移和知识积累，一定会对本能活动产生重大影响。从大量信息中发现知识的活动几乎是伴随着信息的产生而产生，往往伴随着知识推理，如看云识天气。随着信息技术的发展和人类社会的进步，信息量呈现爆炸式增长状态，人工发现知识的能力受到很大限制，因此，我们必须研究智能的知识发现方法。本书主要研究基于学术文献的知识发现问题。基于学术文献的知识发现问题包括：分类、特征提取、术语或实体提取、术语或实体关系提取及本体构建等，其中术语或实体关系提取一般作为领域本体构建的基础工作。由于领域本体收集了领域内相关概念及概念间关系，为基于文献的语义推理提供了可能，因此，领域本体构建工作又为基于文献的知识发现提供了广阔的发展空间。然而，由于科技文献量十分巨大，相关研究还没有有效解决上述问题，因而没有得到大规模推广应用。虽然人们已开展学术论文影响力预测和分类研究，但未将其应用于筛选新出版的领域相关科技论文以便减少文献干扰。现在，学术界纷纷将机器学习方法应用于文献特征提取，在一定程度上获得了良好的效果。但事实上，由于科技论文固有的特性，存在更经济的方法提取文献特征及核心词汇，或构建语义配价模板及例句。

基于科技文献的知识发现主要目标包括以下几个层次：提高标引质量和检索相关性，揭示文献知识点之间的关联线索，挖掘文献中的隐含知识。这些目标的成功实现离不开文献的自动语义标引，这样才能有效地挖掘分散存放且没有直观联系的文献中隐含的知识。相关研究还没有有效解决上述问题，大数据时代的到来使得其实现难度增大。语义知识的缺乏是自动标引的难点之一，比较有效的措施是构建领域本体。研究表明框架语义网络本体(FrameNet)在结构和内容上具有简单可行的推理机制，更适合应用于基于文献的知识发现领域。近年来，基于文献的知识发现研究存在以下问题：虽然确立了领域本体的重要地位，但实际大规模的领域本体构建工作尤其是汉语框架网络语义本体工作还远远不够，基于本体语义推理的知识发现机制还不够完善。

已有的知识发现相关书籍虽然有些也涉及本体构建，但我们认为：框架语义网络本体相比其他本体模型更能支持语义推理，且应建立基于文献评价的本体构建框架。本书介绍

的本体构建工程技术重点解决了以下问题：本体构建基础语料的评价机制，包括学术影响力评价及语料的领域相关性评价、框架核心词筛选测度等，对其他相关文献以及涉及的本体构建的步骤、评价标准、工具等内容不再赘述。采用支持向量机建立学术文献影响力预测模型并分析该模型的学科适用性，作者关键词对大规模文献自动分类的有效性问题，句子向量用于无语义相似性的相同领域文本分类问题，领域框架语义网络本体的词元的领域专指性测度问题，选择框架核心词的策略问题，基于框架词元进行例句的语义角色标引问题。最后以导电聚合物领域本体构建为例演示了本书的本体构建工程技术。

此外，已有的同类书籍往往侧重算法的推导，鲜有配套算法的实现代码，对于初学者而言显得过于艰深。本书论述逻辑严密、条理清楚，所设计的算法均比较简洁，且所有的算法都配套相应的实验和程序代码，更适合作为知识发现领域相关专业的教材或参考书。

本书共包括 7 章，遵循提出问题、分析和解决问题、最终给出结论的研究思路进行组织。

第 1 章，分析研究现状并提出问题。通过对已有研究的分析与评述，确定了本书的内容。

第 2 章，介绍新论文影响力预测问题。讨论基于 SCI 收录的学术文献的结构化信息特征的基于高斯核支持向量机的预测模型的有效性。

第 3 章，介绍术语聚类策略。讨论基于前向、后向及任意位置字符串匹配算法的有效性。

第 4 章，介绍基于词特征向量及句子特征向量单分类器的多分类方法，以 CSCD 数据库收录的中文学术论文为例，比较基于学术文献作者关键词、文本分词、句子向量的文本分类特征选择方法的分类效果。

第 5 章，介绍基于学术文献关键词的汉语领域框架语义网络本体候选核心词词库构建方法及实验。

第 6 章，介绍领域框架语义网络本体库例句库构建方法及语义角色标注方法。

第 7 章，以导电塑料为例，演示基于领域核心词概念的学术文献构建领域本体的框架和效果。

第 1 章至第 7 章由吉久明撰写，第 4 章、第 5 章中使用的中科院分词系统基于 SQL Server 的 ICTCLAS 批处理程序（见附录）由李楠编写。

本书撰写过程中，作者参阅和引用了许多相关研究文献，这为本书的完成提供了帮助，也丰富了本书的素材，在此对相关作者表示真诚的谢意！华东理工大学信息学院高大启教授、华东理工大学科技信息研究所孙济庆研究馆员对本书的撰写提出了宝贵意见，华东理工大学信息学院研究生范琦、中国工商银行北京总行郑荣庭经理对本书使用的科技文献数据处理提供了帮助，在此表示感谢！

此外，本书试图对已有的本体构建工程提出改进，通过基于高水平论文预测的高水平论文样本筛选、基于类别区分度的高核心度或高核聚度的候选词元筛选以提高本体构建质量和效率。因为是初次尝试，疏漏之处在所难免，敬请使用本书的读者批评指正。

作 者

2016 年 12 月

# 目 录

*Contents*

第 1 章

# 绪　论

## 1.1　研究背景及问题提出

### 1.1.1　研究背景

人类文明发展到今天，仍然有许多未知促使人们不断地探索，半个多世纪以前的问题，如人脑的化学反应机理、有机体是否有意识，至今仍未得到完全解决[1, 2]。人们探求未知的途径主要有阅读、实验或调查，在探求未知的过程中，又产生了大量包含知识的信息，这些信息有的留存在人脑中，有的被输出（口头表达、撰写）并储存在特定的文献载体中，且随着信息技术和社会生活的进步，信息呈爆发式增长。仅中国的科学引文数据库（CSCD）收录的2008—2012 年发表的高水平的有关工程类和化学类的学术论文就达 56 万多篇，平均每年 10 万多篇。而学术论文都是由不同的团队或个人独立或协作完成的，大多以一篇论文或图书著作等非结构的形式存储在纸本、数据库或各种机读文件中。存储在人脑或文献载体上的知识在其被再次认识之前是以隐性知识[3]的形式存在的，而这类知识对于人类（确切地说是团体或个人）更好地认知世界具有非常重要的意义[4]。很久以来，将隐性知识显性化的过程即知识发现过程的研究一直是计算机工程专家的研究热点。

知识发现的主要工作包括：收集整理并保存专家经验知识，建立专家系统，从机构或领域信息中发现知识，也包括针对各种结构、半结构或非结构化的数据挖掘[5—10]。

基于半结构或非结构化文献的数据挖掘工作又被称为基于文献的知识发现（虽然文献是存储信息的载体，但学术界习惯用“基于文献的知识发现”表示“基于文献内容的知识发现”），其主要目标包括以下几个层次：提高标引质量和检索相关性[11]，揭示文献知识点之间的关联线索[12—16]，挖掘文献中的隐含知识[14]。这些目标的成功实现离不开文献的自动语义标引，这样才能有效地挖掘分散存放且没有直观联系的文献中隐含的知识。相关研究还没有有效解决上述问题，大数据时代的到来使得其实现难度加大。语义知识的缺乏是自动标引的难点之一，比较有效的措施是构建领域本体[17, 18]。现有领域本体主要包括两类，一类是 WordNet[19, 20]、VerbNet 或 HowNet（简称为 Word 类本体），另一类是框架语义网络本

体[21—25](简称为 FrameNet),前者主要描述词汇概念及概念间相互关系,后者除描述概念及概念间相互关系外,还定义语义配价模型及例句库。研究表明框架语义网络本体具有简单可行的推理机制,更适合应用于基于文献的知识发现领域。框架语义网络本体的基础工作是选择领域术语核心词,在此基础上,建立领域术语之间的关系、例句库及语义配价模型。

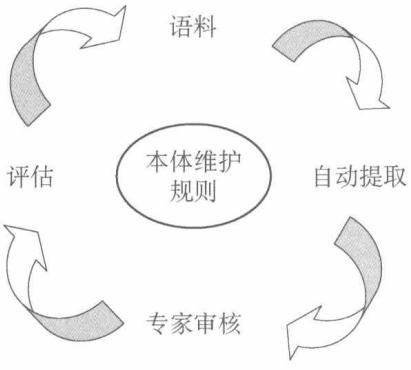


图 1.1 领域本体工程框架

Word 类领域本体工程主要包括以下两个步骤:(1)人工选择或发现领域概念术语,并建立概念术语间关系;(2)收集大量的领域语料,参照已有的领域术语词典(如果有,则主要引用其中的术语概念及其描述,或概念间关系;若没有,则无法参考,直接依赖语料),从语料中提取候选术语及相互间关系,由专家审核其必要性和正确性。其后,再进行本体评估,且遵循本体更新机制进行本体维护,如图 1.1 所示。而 FrameNet 工程则在自动提取阶段增加了例句库及语义配价模板的构建工作。

### 1.1.2 问题的提出

近年来,基于文献的知识发现研究存在以下问题:虽然确立了领域本体的重要地位,但实际大规模的领域本体构建工作尤其是汉语框架网络语义本体工作还远远不够,基于本体语义推理的知识发现机制还不够完善。

具体地讲,已有的本体构建工程存在以下问题:

(1) 缺乏对基础语料的评价机制,包括学术影响力评价及语料的领域相关性评价、框架核心词筛选测度等。已有的学术影响力评价方法不足以解决新发表的学术文献的学术影响力评价问题;文献的领域相关性评价即对文献进行学科分类,大规模文献自动分类较少涉及作者关键词分类有效性研究,虽有研究表明基于句子向量的文本相似度算法效果比基于词向量的文本相似度算法的效果要好[26],但对句子向量用于无语义相似性的相同领域文本的分类研究还很少;有关领域框架语义网络本体的词元收集研究,未涉及其领域专指性测度问题。

(2) 选择框架核心词的策略没有充分发挥学术文献的优势,例句的语义角色标引工作也未充分发挥框架词元的自身优势。

因此,应建立基于文献评价的本体构建框架,如图 1.2 所示。

已有的知识发现研究存在以下问题:由于缺乏足够的领域本体库支撑,有关基于本体语义推理的知识发现的实际应用相当少。

本书拟对框架语义网络本体构建关键问题

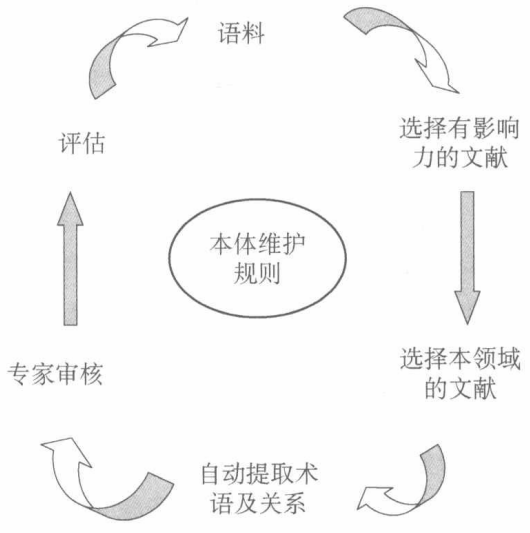


图 1.2 基于文献评价的领域本体工程框架

及基于本体语义推理的知识发现实际应用可行性展开研究，从而为文献知识发现的自动实现提供较好的方法和模型。由于汉语文本的特殊性，已有的技术都受到中文分词、命名实体识别效果的影响，需要进一步研究有效的文本特征选择及例句选择方案。因此本书选择国内高水平学术文献数据库—中国科学引文数据库(CSCD)—文献为样本测试方案的有效性，最后将研究方案应用于导电塑料框架语义网络本体构建及导电塑料制备技术专利文献筛选工作，为专业领域框架语义网络本体的构建提供参考。

## 1.2 国内外研究概况

基于文献的知识发现理论在国内外的研究时间都较长，但国外已开发出如 Arrowsmith 等基于文献内容的知识发现工具[27, 28]，而国内的研究存在一定差距，原因有两个：一是由于国内文献的数字化工作起步较晚；二是因为中英文在语法句法等方面的差异，使得基于中文文献的知识发现存在一定的困难。鉴于支持向量机的优越性能，本书尝试将支持向量机应用于学术文献影响力预测及文本分类。针对文档的特征选择、领域本体库构建的关键步骤(包括：核心词选择、例句语义角色标引及语义配价模板编制)问题的不同特性分别设计相应的解决方案，以实现高效框架语义网路本体库建设的目的。因此，这里将分别介绍基于文献内容的知识发现、支持向量机、文本分类、框架语义网络的国内外研究概况。

### 1.2.1 基于文献内容的知识发现

基于文献内容的知识发现工作常被分为基于相关文献和基于非相关文献的知识发现两种情况：称两篇文献为相关文献即两篇文献关注的主题相同，或存在相同的作者或机构、出版年、出版机构等相同或相近的文献外部特征，或具有相同的参考文献，或被同一篇文献，或同一作者或机构引用等相同或相近的引证特征线索，反之称两篇文献为非相关文献。

“基于相关文献的知识发现”目的在于揭示具有显性相同或相近文献内容特征项文献的不同观点或不同的关注点，揭示具有相同或相近的文献外部特征或引证特征的文献所构成的知识网络或知识图谱。此类知识发现技术主要依赖于统计指标体系为主的词汇共现分析、共同引用或被引分析、共作者(机构)、共出版物等及各种指标的组合分析[29—33]。最近，Liang Wei 等[34]基于上述指标的知识网络空间模型，研究知识的内聚性和知识创作者的领域限制，引入小世界理论将该方向的研究推到一个新的阶段。

“基于非相关文献的知识发现”由 Swanson[4, 28]报道深海鱼油对于雷诺氏病治疗作用的论文中首次提出，使用“非相关文献”的术语，意在区分其研究方法与已有的基于文献的外部显性特征关联线索进行知识发现的研究方法的不同。该项工作基于以下三段式的假设：A 引起 B(A 与 B 相关)，B 引起 C(B 与 C 相关)，则 A 引起 C(A 与 C 相关)。Swanson 将世界分为三个层次：物理的世界(World 1)；基于心理状态或过程的经验世界(Word 2)；基于问题、理论或其他人类思维的产物等客观知识世界(World 3)。第三层次的世界是由人创造的，但其中存在许多远未认知的知识。提出基于信息检索的非相关文献的知识发现方法旨在揭示更多不为人知的但已经存在的知识，这些知识之所以未被发现，是因为人类索引及获

取信息能力的限制。

Swanson[28, 35]设计的基于非相关文献的知识发现模型包括开放式和封闭式两种：

开放式模型主要在科学假设形成阶段，选择感兴趣的课题，意在找到与该课题有关的隐含知识。以某一主题词 I 检索得到相关文献集合 A，提取与词 I 共同出现在题目中的术语词汇形成词汇表 B，以 B 中的词分别检索得到相关文献集合，提取与 B 词汇表中的词共同出现在题目中的术语词汇形成词汇表 C，进一步借助统计分析方法，可以建立主题词 I 经由词汇表 B 中的某些词与词汇表 C 中的某些词概念相关的假设知识集合 H，如图 1.3 所示。

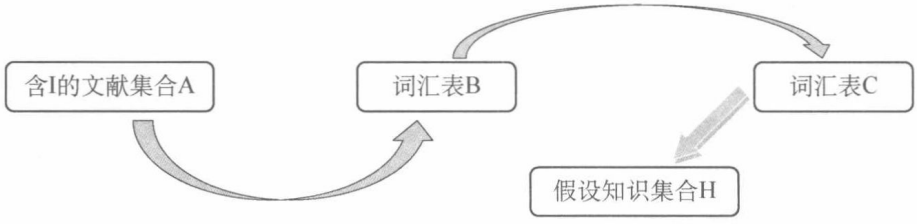


**图 1.3 开放式的科学假设形成框架**

封闭式模型主要用于科学假设的验证阶段，选择假设知识中的两种主题词 Ia 和 Ic，分别检索获得相关文献集合 Ba 和 Bc，并从 Ba 与 Bc 中分别提取与 Ia 和 Ic 共同出现在同一篇文献中的术语词汇形成词汇集合 B，进一步借助统计分析方法，可以为科学假设建立经由主题词 Ia 和 Ic 找到的词汇集合 B 中的某些词汇所提供的证据候选集合 V，如图 1.4 所示。


**图 1.4 封闭式的科学假设验证框架**

之后有学者相继对 Swanson 的模型进行了修改，分别提出了基于关键词汇组成的短语词频[36]、基于概念[37]、基于概念词频[38]、基于领域本体[39, 40]的开放式或封闭式知识发现模型，也有研究探讨选择文献中不同位置的词或概念、集合 B 中词汇的各种不同选择策略[41]对知识发现效果的影响。但基于领域本体的知识发现模型[39, 40]实际上是对基于概念词频等所谓的语义关联挖掘的简单推广，未见基于框架语义网络本体的知识发现模型。

冷伏海团队[42]首次将 Swanson 等人的研究介绍到国内，在国家自然科学基金的资助下针对 B 词集的选择开展了大量的研究。遗憾的是该团队的工作对国内的基于非相关文献的隐含知识发现研究的推动不是太大。尽管有人基于 Swanson 方法开展了某些领域的实证研究[43, 44]，至今仍未出现成熟的可以与 Arrosmith 系统媲美的基于汉语文献的隐含知识发现系统[45]，对基于领域本体[39, 42]的知识发现模型的后续研究主要转向了基于语义的期刊出版研究[46, 47]。

我们认为，领域本体构建已经成为各项基于语义应用发展的瓶颈，应加紧开发领域本体库，尤其应该加紧研究领域框架语义网络本体库的建设，并研究基于框架语义网络本体的文

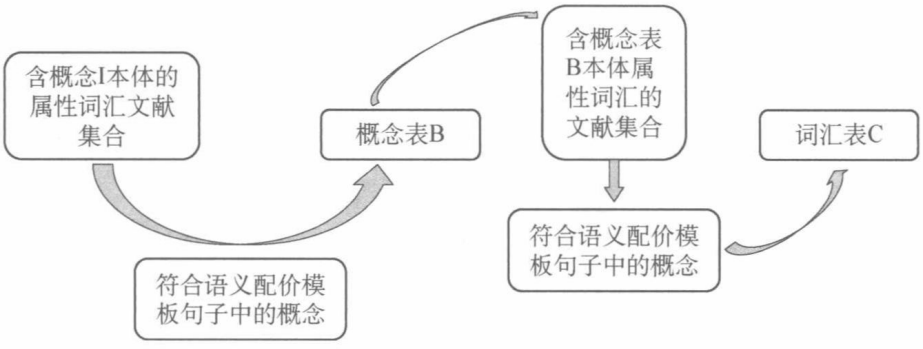


**图 1.5 基于框架语义网络本体的开放式知识发现模型**

献中隐含知识发现的模型，如图 1.5 所示，由于框架语义网络提供了语义配价模型和例句库，可以将基于概念的挖掘上升到基于句子包的隐含知识挖掘，预期能取得更好的效果。

### 1.2.2 支持向量机

支持向量机(Support Vector Machine, SVM)[48—54]是在 20 世纪 90 年代发展起来的一种基于结构风险最小化准则的分类学习机模型，是由贝尔实验室的 Corinna Cortes 和 Vladimir Vapnik 受 Fisher 算法和神经网络感知器算法的启发于 1993 年创建[81]，Fisher 算法通过寻找两类问题分割面，神经网络感知器将非线性空间映射为线性空间。它通过构造并求解目标函数来获得两类样本数据之间的决策超平面，以保证最小的分类错误率。这一新兴的学习机模型已经在手写数字识别、三维目标识别、人脸识别、文本图像分类、时间序列预测、主成分分析、生物标志、水域物种风险判别等实际问题的应用中，表现出了良好的分类或识别能力。从实际分类效果来看，支持向量机在解决小样本、非线性及高维的模式识别问题方面是目前已知的分类器中效果表现较优的一种机器学习方法。如今，支持向量机及其应用研究已引起越来越多的兴趣和关注，成为机器学习理论和技术领域中的一个新热点。

支持向量机将输入空间的向量按事先选定的非线性映射映射到某个高维的特征空间。在这个空间，一个线性判决面是用特殊的能够保证网络具有高推广能力的特征构建的，如图 1.6 所示[50]。

图 1.6 表明，构建一个最优的超平面只需要考虑较少的训练样本数据，这些样本向量称为支持向量，它们决定了两类之间的最大间隔。若训练集样本被一个最优超平面完全正确地区分开，那么测试集误差概率的期望值就小于支持向量个数的期望值与训练样本个数的比值。即：

$$E[\Pr(error)] \leqslant \frac{E[number\ of\ support\ vectors]}{number\ of\ training\ vectors} \tag{1-1}$$

这个上界并没有涉及特征空间的维数，这表明如果能够从训练集找到少量的几个支持向量构建一个最优的分界超平面，即使输入空间是无限维的，超平面的推广能力也会很好。

用支持向量机进行分类必然有两种情况：一是所求得的支持向量机能够完全区分两类样本，二是在区分两类样本时存在误差。

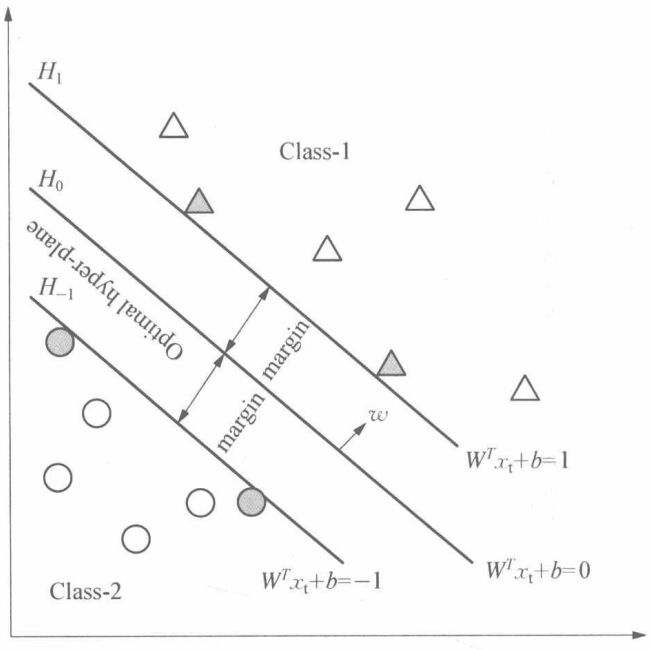


**图 1.6 支持向量机模型示意图**

设训练样本集为：

$$(y_1, x_1), \cdots, (y_l, x_l), y_1 \in \{-1, 1\} \tag{1-2}$$

称样本集为线性可分的，即：存在向量 $\omega$ 和 $b$，使得：

$$\omega \cdot x_i + b \geqslant 1 \quad \text{if} \quad y_i = 1 \tag{1-3}$$

$$\omega \cdot x_i + b \leqslant 1 \quad \text{if} \quad y_i = -1 \tag{1-4}$$

上式可以改写为：

$$y_i(\omega \cdot x_i + b) \geqslant 1, \ i = 1, \cdots, l, \ y_i \in \{-1, 1\} \tag{1-5}$$

使得 $y_i(\omega \cdot x_i + b) = 1$ 的向量 $x_i^0$ 被称为支持向量，此时，

$$\omega = \sum_{i=1}^{l} y_i \alpha_i^0 x_i^0, \text{其中 } \alpha_i^0 \geqslant 0 \tag{1-6}$$

寻找支持向量的问题转化为选择适当的核函数和参数，解决以下最优化问题：

$$\min \varphi(\omega) = \frac{1}{2} \| \omega \|^2 = \frac{1}{2} \omega^T \omega$$

$$s.t. \quad y_i[\omega^T x_i + b] - 1 \geqslant 0 \quad i = 1, 2, \cdots, N \tag{1-7}$$

当训练样本不能被完全区分时，将决策函数调整为：

$$y_i(\omega \cdot x_i + b) \geqslant 1 - \xi_i, \ i = 1, \cdots, l$$

$$\xi_i \geqslant 0, \ i = 1, \cdots, l \tag{1-8}$$

去掉那些被错分的样本，其余样本将被完全区分。其支持向量选择问题转化为选择适当的参数，解决以下最优化问题：

$$\min\left(\frac{1}{2}\|\omega\|^2+C\sum_i^m\xi_i\right)$$
$$s.t. \quad y_i[\omega\cdot x_i+b]-1+\xi_i\geqslant 0$$
$$\xi_i\geqslant 0 \tag{1-9}$$

求解该问题的常用方法为拉格朗日乘子法，便将上述问题转化为求解下述优化问题：

$$\max\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_i^l\sum_{j=1}^{l}a_ia_jy_iy_j(x_i\cdot x_j) \tag{1-10}$$

$$s.t. \quad \sum_{i=1}^{l}a_iy_i=0 \tag{1-11}$$

$$0\leqslant a_i\leqslant C,\ i=1,\cdots,l \tag{1-12}$$

这里，$a_i$ 为拉格朗日乘子，上述问题的解中 $a_i\neq 0$ 所对应的样本即为支持向量。

对于非线性可分问题，将样本空间映射到线性可分的特征空间。这种经过映射后再求分界面的过程，已经被证明可以用符合 Mercer 核条件的核函数方法完成，避开了高维空间内的点积运算[68]。

常用的基本核函数有：

$$K(u,\ v)=\exp\left(-\frac{|u-v|}{\sigma}\right) \tag{1-13}$$

多项式核：

$$K(u,\ v)=(u\cdot v+1)^d \tag{1-14}$$

高斯核：

$$K(u,\ v)=\exp\left\{-\frac{|u-v|^2}{\sigma^2}\right\} \tag{1-15}$$

支持向量机具有全局性、简单性、分类正确率高等很重要的特点，成为许多领域内解决分类问题的首选方法。但基本的支持向量机在实际应用中遇到了各种问题，包括核函数的选择或构造、不可分问题、训练效率问题等。研究人员主要从两个方面来解决这些问题：一是将支持向量机与其他算法相结合，克服支持向量机的缺点；二是研究支持向量机的特征选择算法[55—57]。具体地有以下几个主要研究方向：

(1) 核函数及参数选择

研究人员已经构造了许多类型的核函数，如：多项式核 $k(u,\ v)=(u\cdot v+c)^d$ 及其改进核[58]，Gauss 核等，还有其他的核函数，如：B-样条核、傅立叶核、Sigmoid 核等常用经验核[59]，Kuo，Bor-Chen[60]、Motai，Yuichi[61]、Cristianini[62] 和 Evgeniou[63]、Haussler[64]、Bernhard[65]、Watkins[66]、Jaakkola[67] 等则是直接从数据中学习，从而构造或改进核，邓乃

扬等[68]系统地介绍了构造核函数的原则，Amari S[59]等研究了核函数的构造技巧。针对以RBF函数为核函数的支持向量机，不少学者研究了核函数宽度参数的优化情况，通常利用LOO方法和K-Fold交叉验证方法估计模型的推广能力，由于这两种方法计算量较大，又有采用Maximal-Discrepancy准则、Radius/Margin Bound方法、导数平方和准则、三步搜索技术、网格搜索技术、遗传算法、GA及Bootstrap方法进行参数优化的研究[69—73]。

也有少部分的学者研究无参数化的支持向量机[74—76]。其基本思想是减少常用的惩罚参数，如：文献[76]中，将基本支持向量机问题转化为式(1-16)的最优化问题，只要选择适当的核函数即可。

$$\begin{aligned} &\min \quad e'\xi \\ s.t. \quad &(K(A, A')u+eb) \leqslant \xi, \\ &y-(K(A, A')u+eb) \leqslant \xi, \end{aligned} \tag{1-16}$$

另外，适合字符串、基因序列等结构化数据的卷积核也被开发出来，使用卷积核将相关的特征项合并减少向量的维数，进而有利于提高算法的性能[77—80]。

(2) 不可分问题

不可分问题是分类研究中的重要问题之一，自然也是支持向量机必须面对的。目前被广泛使用的解决方案是将模糊数学与支持向量机结合的模糊支持向量机，该方法在Vaplik[81]不可分类方法的基础上，通过各类内样本对所属类别的隶属度来寻找特征向量，其基本思想如下[50]：

设：

$$(y_1, x_1, s_1), \cdots, (y_l, x_l, s_l) \tag{1-17}$$

为给定的样本，其中 $x_i$ 为 $n$ 维的训练样本向量，分别以 $s_i$ 的模糊隶属度属于类别 $y_i$，$y_i$ 为1或$-1$，$\sigma \leqslant s_i \leqslant 1$，$\sigma$ 足够小且大于0。最优超平面为以下问题的解：

$$\begin{aligned} &\min \quad \frac{1}{2}\omega \cdot \omega + C\sum_i^l s_i\xi_i \\ &y_i(\omega \cdot x_i + b) \geqslant 1-\xi_i, \ i=1, \cdots l \\ &\xi_i \geqslant 0, \ i=1, \cdots, l \end{aligned} \tag{1-18}$$

式(1-18)与式(1-9)的区别在于前者赋予不同样本不同的隶属度权重。实际应用中，不同的样本隶属度计算方法对支持向量机分类算法的正确率有很大的影响。Morikawa, Kazuya[82]、Wu, Zhenning[83]针对传统支持向量机中存在对噪声或野值敏感的问题，提出了一种基于紧密度的模糊隶属度，即：

$$\mu(x_i) = \begin{cases} 0.6 \times \left(\dfrac{1-d(x_i)/R}{1+d(x_i)/R}\right) + 0.4, & d(x_i) \leqslant R \\ 0.4 \times \left(\dfrac{1}{1+(d(x_i)-R)}\right), & d(x_i) > R \end{cases} \tag{1-19}$$

其中，$R$ 为样本集中最小包围球半径；$d(x_i)$ 为样本集中样本 $x_i$ 与其最小包围球中心 $a$ 之间