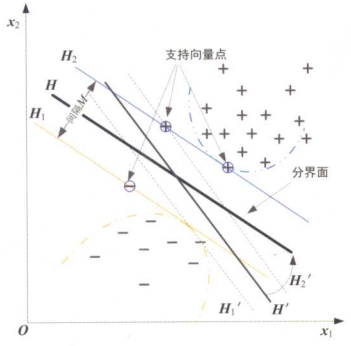
创新技术学术专著

INNOVATIVE

基于支持向量机的聚类及文本分类关键技术研究

Research on Key Techniques of Clustering and Text Categorization based on Support Vector Machines

源 周亚建 杨义先 著



人民邮电出版社
POSTS & TELECOM PRESS

创新技术学术专著

INNOVATIVE

基于支持向量机的聚类及文本分类

关键技术研究

Research on Key Techniques of Clustering and Text Categorization based on Support Vector Machines

平源 周亚建 杨义先 著

人民邮电出版社

北京

图书在版编目（CIP）数据

基于支持向量机的聚类及文本分类关键技术研究 / 平源，周亚建，杨义先著. -- 北京 : 人民邮电出版社，2014.1
ISBN 978-7-115-33269-1

Ⅰ. ①基… Ⅱ. ①平… ②周… ③杨… Ⅲ. ①信息泄漏－研究 Ⅳ. ①TN918

中国版本图书馆CIP数据核字(2013)第251313号

内 容 提 要

本书在国内外已有相关工作成果的基础上，研究了以支持向量机理论为基础的聚类分析方法及文本分类中的关键技术，内容涉及模型基本原理、参数分析、数据预处理、聚类分析关键问题与改进及文本表示策略等，旨在系统地总结作者近年来在该领域的研究工作，并帮助读者快速了解影响支持向量聚类和分类器的关键因素及发挥优势、规避局限的分析方法和改进思路。

本书可供从事机器学习、文本分类相关理论及应用研究，尤其对支持向量机相关研究感兴趣的科研、教学和工程技术人员参考。

◆ 著　　平　源　周亚建　杨义先
责任编辑　代晓丽
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷

◆ 开本：700×1000　1/16
印张：12　　2014 年 1 月第 1 版
字数：228 千字　　2014 年 1 月北京第 1 次印刷

定价：68.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前　言

随着大数据时代的来临，互联网上分布、流动并急剧膨胀的不仅有多样化应用所产生的具有可用性、有效性的内容资源，还充斥着大量干扰正常业务、侵犯隐私、误导公众甚至危害社会稳定并同样多样化的信息和行为。从数据管理的角度，有必要根据不同行业、领域用户的需要，快速、高效地组织、分析、提取并分级保护有用的数据或敏感信息；从内容安全的角度，人们更期待能够对正在或即将泄露的敏感信息进行检测和保护，对存在虚假、恶意或诱导意图的内容或行为进行分类、过滤和分析以及时发现攻击源、保护受害者，并调动智能防御系统进行数据处理、知识学习和模型更新。在众多机器学习方法中，聚类分析和分类被认为是快速、准确地发现、定位、组织和分析具有特定用途的可用信息和行为模式，实现信息安全保护效率最大化的有效途径和关键技术。

作为一种基于统计学习理论的机器学习方法，支持向量机不仅具有优秀的小样本学习能力，而且较好地解决了非线性、高维度、局部极小值等问题。它既能通过构造闭合分界面来进行无监督的数据聚类分析，又可以通过构造非闭合分界面来处理有监督的数据分类问题，尤其适于处理高维、稀疏且特征之间具有较大相关性的文本数据，因而具有高效地解决前述以数据管理和内容安全为目的数据分析问题的优秀品质。然而，当样本规模较大、维数较高、类别数较多且存在噪声数据干扰时，传统的基于支持向量机的聚类分析模型存在训练速度较慢、参数敏感且难以找到合适的簇原型来提升簇标定的效率和准确率，以至于其对任意簇形状数据描述能力的优势难以得到发挥，甚至出现了应用研究和发展的瓶颈；作为互联网信息存在的主要形式，文本数据通常具有前述特征，并且会以降低数据可分性的方式影响基于支持向量机的文本分类系统性能，包括降低训练和分类速度、准确率以及收集到的支持向量样本的指示意义等。因此，要解决这些问题，不仅要针对性破解聚类/分类器模型瓶颈，而且需要从待分析对象的角度寻求优化表示以助于发挥模型的优势。

本书凝聚了笔者近年来在聚类分析和文本分类领域的科研经验与思考，得到了国家高技术研究发展计划（“863”计划）“面向敏感用户数据防泄露的统一模型与方法（2009AA01Z430）”和国家自然科学基金（60972077，61121061，61161140320）等项目的支持，也包含了与田英杰研究员、彭建芬、彭维平、郭春博士和薛超、薛凯、李正、程丽硕士一起合作的部分研究成果，在此一并表示感谢。

由于作者水平有限，本书中难免有错误或者不周之处，敬请广大读者批评指正。

作者

目　录

第 1 章　绪论……1

1.1　引言……1

1.2　机器学习理论 ……2

1.2.1　无监督学习 ……3

1.2.2　有监督学习 ……3

1.2.3　半监督学习 ……3

1.2.4　增强学习 ……4

1.3　支持向量机与聚类分析 ….4

1.4　支持向量机与文本分类 ….7

1.5　本书的主要工作 ……10

第 2 章　支持向量机技术基础 ……13

2.1　引言……13

2.2　统计学习理论 ……13

2.3　支持向量机技术 ……16

2.3.1　支持向量分类机 ……16

2.3.2　L2-支持向量机……19

2.3.3　多类问题的决策方法……21

2.3.4　支持向量回归机模型……23

2.3.5　支持向量机研究现状……24

2.4　支持向量聚类……29

2.4.1　支持向量聚类模型…29

2.4.2　影响支持向量聚类的关键因素……32

2.5　本章小结……40

第 3 章　双质心支持向量聚类……41

3.1　引言……41

3.2　噪声数据点消除策略……43

3.2.1　噪声数据分布结构分析……43

3.2.2　噪声数据消除算法……45

3.3　双质心簇标定策略……47

3.3.1　簇的分解策略……47

3.3.2　单组件双质心的构造……48

3.3.3　成员关系的判定规则……50

3.3.4　算法描述……51

3.4　DBC 时间性能分析……52

3.5　聚类实验分析……53

3.5.1　数据集……53

3.5.2 实验对比算法 ……53
3.5.3 噪声数据消除实验 …54
3.5.4 DBC 聚类效果测试 …57
3.5.5 DBC 整体性能测试 …59
3.5.6 DBC 模型的半监督应用测试……60
3.6 本章小结……60
第 4 章 基于凸分解的簇标定算法……63
4.1 引言……63
4.2 基于凸分解的簇标定算法 ……64
4.2.1 簇在特征空间中的凸性质……64
4.2.2 支持超凸多面体的凸分解……65
4.2.3 凸包的标定算法 ……72
4.2.4 标定非凸包样本 ……76
4.3 CDCL 算法时间性能分析 ……76
4.4 聚类实验分析……78
4.4.1 数据集……78
4.4.2 实验对比算法 ……79
4.4.3 CDCL 算法适应能力分析……80
4.4.4 CDCL 算法整体性能测试……82
4.5 本章小结……87
第 5 章 快速支持向量聚类算法… 89
5.1 引言……89
5.2 快速支持向量聚类算法（FASVC）……90
5.2.1 选择簇边界样本……90
5.2.2 构造超球面……92
5.2.3 自适应的簇标定策略……96
5.2.4 FASVC 算法的实现……97
5.3 FASVC 时间性能及特点分析……99
5.3.1 FASVC 时间性能分析……99
5.3.2 FASVC 算法特点 … 100
5.4 聚类实验分析…… 101
5.4.1 数据集…… 102
5.4.2 实验对比算法…… 102
5.4.3 FASVC 参数敏感性测试…… 103
5.4.4 FASVC 算法整体性能测试…… 104
5.4.5 利用 FASVC 进行文本聚类…… 110
5.4.6 利用 FASVC 识别 P2P 流量……111

5.5 本章小结……112
第6章 基于支持向量机的多模式文本分类研究……113
6.1 引言……113
6.2 文本表示的关键问题与启示……114
6.2.1 场景1：特征的文档频率之外的信息……116
6.2.2 场景2：最大值保留的特征权重与特征的多类别分布信息……116
6.2.3 场景3：文本的结构信息……118
6.3 基于支持向量机的多模式文本分类方案……120
6.3.1 自适应的文本块划分算法……120
6.3.2 兼顾类别贡献度和类间区分度的特征权重方案……121
6.3.3 融合多类别倾向的特征类间区分能力强化方案……122
6.3.4 基于文本块重要性分布加权的特征频率方案……124
6.4 分类实验分析……125
6.4.1 数据集……125
6.4.2 实验对比方案……127
6.4.3 评价指标……129
6.4.4 CCE方案实验结果与分析……130
6.4.5 C2TCTVT算法框架实验结果与分析……135
6.4.6 NWET与N2WET组合方案实验结果与分析……145
6.5 本章小结……153
结束语……154
参考文献……158
名词索引……180

第1章 绪 论

1.1 引言

随着信息技术的发展，互联网上分布和流动的内容资源呈现海量和多元化的膨胀之势，同时，得益于数据收集和存储技术的快速进步使得各组织机构可迅速积累大量的数据。这些以不同的状态（如静态、传输态及使用态等）呈现的数据中，既有多元应用产生的正常可用的数据，如新闻、电子邮件、新闻组、微博、即时通信、P2P 流量、安全策略、系统日志等互联网（或系统）内容资源及金融、市场、医学、教育、政府等行业数据，也有虚假信息、钓鱼网站、色情网站、垃圾邮件、欺诈广告、个人隐私、恶意软件、僵尸网络攻击流、网络经济犯罪等影响信息或资源可用性、泄露敏感信息、误导公众、危害社会稳定、涉及国家重大利益等内容。对于前者，从信息的可用性、有效性的角度，有必要根据不同行业、领域用户的需要，快速、高效地组织、分析、提取和分级保护有用的数据；而对于后者，则无论从真实性、可用性、道德性、合法性，还是内容安全的角度，都更有必要对正在或即将泄露的敏感数据进行检测和保护，对存在虚假、恶意或诱导的内容或行为进行分类、过滤和分析，以便及时定位受害者或攻击源，同时调动智能处理系统进行数据处理、知识学习。然而，数据并不等于知识或者信息的事实导致：一方面，分布于全球范围处于静态存储、传输或使用态的海量数据，其无论从时间还是空间的角度都给人们分析、理解和决策带来极大的难度；另一方面，即便是一些相对较少的内容或行为数据集，也可能因为其数据本身的一些非传统特点，使人们在依据特定应用目的或安全策略进行定位敏感信息、分析网络行为和提取可用知识等工作时困难重重。但是，在信息资源高度数字化和网络化的今天，类似于“维基解密”、“Facebook 会员资料外泄”及“高盛资料内部窃取”等敏感信息泄露和 Sybil 攻击[1]、高级持续性渗透（Advanced Persistent Threat，APT）[2]等网络攻击却频繁出现，因此，无论是从数据管理还是从内容安全的角度，都迫切需要在发生各类信息安全事件之前，快速、自动地发现、定位和组织具有特定用途的“可用”信息和行为模式，以使

得信息安全保护效率最大化。

数据挖掘将传统的数据分析方法建立在能够集中或分布处理大规模数据的优秀算法之上，具有通过对大型的数据存储结构中的每个数据进行分析，找到大规模数据之间的某种规律，进而提取出符合领域需要的有用信息的能力[3]。作为数据挖掘的两个关键研究领域，聚类分析和分类为面向信息安全为目的的特定信息发现、提取和目标行为模式分析等提供了有效的解决途径。

聚类分析是将包括对象、数据或特征向量在内的模式（Pattern）以非监督的方式划分到不同簇类的过程[4,5]。其目的是通过某种相似测度（如欧式距离、马哈拉诺比斯距离[6]、余弦等）发现存在紧密关系的观测值簇，使得簇内部的对象彼此之间的相似度尽可能地大，而不同簇类的对象之间的相似度尽可能地小，甚至不同或不相关。分类的任务则是通过一定数量的样本或实例学习后，建立一个分类模型（或带约束目标函数），利用它将新的对象映射到一个预先定义的类别标号。由于聚类分析能为探索未知的数据结构提供帮助，其分析结果可为分类等任务提供优选的样本集，降低人工标注等经验分析的工作量，因而成为一系列数据分析的起点；而分类则为分析已知的数据结构提供帮助，并被广泛用于描述性建模和预测性建模。二者的结合，使得数据挖掘技术不仅可被用于在大量数据中发现未知的有用模式，并且具有预测未来观察结果的能力，并广泛用于分类特征和数据集优化[7,8]、入侵检测[9~12]、安全日志分析[13]、网络流量与行为分析[11,14]、隐私保护的数据分析[15]、安全过程控制[16]、风险评估[17]、网页分析与垃圾信息清理[18~20]、网络欺诈检测[21]、系统漏洞挖掘[22]、社会网络分析[23,24]、信息检索与文本摘要、生物信息等科研领域或实际应用中。

通过对敏感数据保护、网络流行为分析等相关领域的研究和数据分析发现：

（1）不同背景和应用的数据或特征通常表现出不同的分布结构，这对聚类分析发现任意形状簇类的能力提出了新的要求。

（2）互联网及各行业领域的信息均不同程度地表现为文本形式（包括 Web 3.0 中的异构资源语义标注），以文本为对象的数据分析与分类不仅是数据挖掘和信息检索的重要基础，也仍然是内容安全领域的关键内容。更重要的是，从建立学习模型的角度，聚类分析和分类同属于机器学习的范畴，因此，从机器学习的角度研究更为有效地解决聚类分析对任意簇形状的处理和大规模的文本分类有着重要的理论意义和应用价值。

1.2 机器学习理论

机器学习是数据挖掘的三大支柱之一。Tom M. Mitchell 认为机器学习就是“计算机利用经验改善系统自身性能的行为”[25]。因此机器学习的研究成果是支撑数据

挖掘发现未知有用模式和预测未来观察结果的能力的关键。例如，在敏感数据保护中，要定位敏感信息，一般需要通过对一定数量包含有敏感词汇的样本进行学习，建立正常和敏感信息模式，然后利用这个模式来扫描待保护区域，发现并保护包含敏感信息的载体。一般而言，根据计算机可利用“经验”的差异，可将机器学习分为无监督学习（Unsupervised Learning）、有监督学习（Supervised Learning）、半监督学习（Semi-supervised Learning）和增强学习（Reinforcement Learning）。

1.2.1 无监督学习

任何学习过程，都有训练集的给定，而无监督学习过程的训练集不包含人工标注。例如，对于给定的训练集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，其中，$\boldsymbol{x}_i\in\boldsymbol{R}^d(i=1,\cdots,N)$ 为 d 维样本。由于仅能得到一个没有任何指示信息的 $N\times d$ 训练矩阵，无监督学习的目的就是要探索训练集中蕴含的某种分布结构信息，发现训练集中样本之间的某种关系或规律。例如在流量分析中，无监督学习可用于发现流量的分布规律，而这一规律可用于帮助对主机的 P2P 行为的识别。无监督学习主要包含聚类分析、维度规约和孤异点检测等内容[26]。

1.2.2 有监督学习

与无监督学习不同，有监督学习通过分析具有“输入—输出”模式的数据，寻找某种逼近输入“$\boldsymbol{X}$”和输出“$\boldsymbol{y}$”之间的关系函数 f。对给定输入 $\boldsymbol{X}$，利用该函数可得到具有最小误差的输出 $\boldsymbol{y}$。设训练集 $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}\in\boldsymbol{X}\times\boldsymbol{y}$，且 $\boldsymbol{X}\subset\boldsymbol{R}^d$。则有监督学习过程则是通过对训练集的分析，建立

$$y_i=f(\boldsymbol{x}_i|\boldsymbol{\Theta}),\quad i=1,\cdots,N \tag{1-1}$$

其中，$\boldsymbol{\Theta}$ 为参数集，通过优化 $\boldsymbol{\Theta}$ 使得 $f(\bullet)$ 在预测输入输出关系时，误差尽可能地小。当输出 $\boldsymbol{y}$ 为具体的类别标记时，该学习模型属于有分类（Classification）问题，$f(\bullet)$ 称为决策函数；否则属于回归（Regression）问题，$f(\bullet)$ 称为回归函数。由于监督学习过程有指示信息，所建立的分类或回归模型往往有较高的准确率，因此得到广泛的应用，本书第 6 章将重点研究其在文本分类领域的应用。

1.2.3 半监督学习

由于有监督学习在进行模型训练时往往需要大量人工标记的标签数据（Labeleddata），而要得到这样的完备数据集非常困难。例如在进行垃圾邮件过滤、色情网页检测分析时，几乎不可能获取到符合异常信息的所有模式，但却

容易获取大量无标签数据（Unlabeleddata）。此时就可使用半监督学习方法，它介于无监督和有监督学习之间，由于已有少量指示信息可利用，一定程度避免了无监督学习的相对盲目性，又利用了无监督的某种相似测度而大大降低了人工标注的工作量。因此，在使用半监督学习方法时，训练集中既有带标记数据，又有无标签数据同时参与建立学习模型。

1.2.4 增强学习

增强学习源于控制理论，它是一种从环境到行为映射的学习模式，目标是使奖励函数值最大，即所谓的“状态—动作—奖励”（State-Action-Rewards）三元组[26]。增强学习不同于有监督学习的最重要特征是，有监督学习需要通过标签数据去告知学习系统如何产生正确的动作或结果，而是由环境提供的对产生的动作的好坏给出一个标量的评价。所以，增强学习的问题是去学习如何将条件映射为具体的动作才能最大化奖励或最小化惩罚。

1.3 支持向量机与聚类分析

前面提到，聚类分析实际上是将无标记的样本按照某种相似性度量进行分组（成簇），而这种相似性的度量一般采用某种距离测度来反映，如欧式距离、余弦距离、曼哈坦距离、明考斯基距离、相关系数[27]及马氏距离等[28, 29]。其中，欧式距离在聚类分析中应用最为广泛。经过近50年的研究，形成了一系列重要的聚类分析方法。

（1）基于划分的方法（Partitioning Method）

对于给定的一个含有 N 个对象的数据集，该方法基于某种距离测度寻找对数据集的 $K(K<N)$ 个划分（簇），该划分通常以簇类相似度与簇间相似度的最大比值为停止条件。其主要算法包括 K 均值（K-means）算法[30]、K 中心点（K-medoids）算法[31]和 CLARANS 算法[32]等。由于该类算法受 K 个初选样本的影响较大而不太稳定，并且得到的是对数据输入空间的球状划分[33,34]，因此出现了一些仅使用核心样本进行划分的改进方案，如 FEKM（Fastexact K-means）[35]和模糊 K 均值聚类（Fuzzy K-means）[36]等。

（2）基于层次聚类的方法（Hierarchical Method）

基于划分的方法需要对 K 值进行人工设定，对于簇个数未知而需要探索的情况无能为力，而基于层次的聚类为此提供了有效的解决途径。该方法对给定的数据集进行层次分解。若采用自下而上，先将每个数据对象形成单独的组，再逐一合并最近邻的数据对象或组，直到满足某个终止条件，称为“聚合（Agglomerative）”法；反之，采用自上而下，先将所有数据对象归为一个簇中，再迭代分裂为更小的簇的方法称为“分裂（Divisive）法”。其典型代表分别为

BIRCH[37]和 CURE 算法[38]。然而基于层次聚类并非某种算法所独有，其他聚类算法同样可以使用该模式探索或优化簇数量，比如文献[39]和文献[40]则是分别采用了聚合法和分裂法的支持向量聚类（Support Vector Clustering，SVC）算法。

（3）基于密度的方法（Density-based Method）

该方法期望簇内对象尽可能的稠密，只要以某个数据对象为中心的一个区域中的样本密度大于某个阈值，就将其归入到近邻的簇中。由于该方法指定了区域内的密度要求，可以一定程度上发现具有不规则形状的簇，主要的算法包括 Rough-DBSCAN[41,42]、OPTICS[43]和 DENCLUE[44]等。通常可以将基于密度的方法与基于层次聚类的方法相结合，将密度阈值作为簇分裂或合并的参考。

（4）基于网格的方法（Grid-based Method）

聚类分析的本质就是根据某种相似原则对数据空间进行划分，而划分得到的轮廓有不同的结构。前面的 K 均值方法得到的是“类圆/球形”轮廓，而基于网格的方法则将数据空间量化并划分为有限个网格单元（Grid Cell），然后以这种单个的网格单元为对象进行聚类分析。这种以网格单元作为聚类的基本单位，而不是单个样本，好处是在处理大规模数据集时可以将数据集预划分（注：这样划分无需用 K 均值方法），然后实现类似于块状的快速聚类，其处理时间只与网格单元数量有关。如果以簇原型（Cluster Prototype）的角度来看，可以认为这是一种将多个网格单元做簇原型使用的方法。其代表算法有 STING[45]、CLIQUE[46]、Wave-Cluster[47]及其变体，另外文献[48]的 SVC 方案中也采用了类似的思路。

（5）模糊聚类（Fuzzy Clustering）

在对处于簇轮廓边缘的对象进行聚类时，可能出现将其归为相邻的任何一个簇都是有意义的情况，而且不同的划分会带来截然不同的结果，因此通常在此引入一个隶属度的概念，即该数据对象同时倾向于多个簇，但对于每个簇的隶属度或者置信度不同，然后根据具体的策略将其指派到某一个簇中。其主要的算法包括模糊 C 均值算法（Fuzzy C-Means）[49]、模糊支持向量聚类算法（Fuzzy Support Vector Clustering，FSVC）[39]等。

（6）基于群智能优化的方法（Swarm Optimization based Method）

该类算法和前述算法并没有本质上的不同，只是每一种算法在探索数据分布结构的时候都需要有相应的约束或停止条件，为此基于群智能优化的簇划分方法则采用一些智能优化算法，如遗传算法（Genetic Algorithm，GA）[50]、粒子群算法（Particle Swarm Optimization，PSO）[51]、蚁群算法（Ant Colony Optimization，ACO）[52,53]以及菌群算法（Bacterial Foraging，BF）[54]等改进搜索策略，降低迭代寻找最优解的复杂度。因此，该类算法并没有改变聚类的划分原则或相似性判别规则。

上述的聚类算法基本上都是针对线性可分的数据空间，因此并不适合处理那

些线性不可分的数据，它们很难得到合适的聚类轮廓和期望的簇数量，并且聚类的效果也不够稳定。因此，针对非线性可分的聚类问题，最新的研究提出通过谱聚类（Spectral Clustering）[55]和基于核函数（Kernel Function）变换[56]的聚类方式。谱聚类的代表算法有浙江大学蔡登教授等[57]设计的基于图规整非负矩阵分解方法（GNMF）的谱聚类以及南京大学周志华教授等[58]设计的多流形谱聚类算法。为了能在任意形状的样本空间上聚类，且收敛于全局最优解，谱聚类方法需要利用所有样本点的相似关系建立矩阵，然后获取矩阵的前 n 个特征向量来聚类不同数据点，其除了求解特征值与特征向量需要较大的时间复杂度外，还存在需要人为指定簇的数量和聚类效果的顽健性不高等局限，并且一个数据集上聚类的结果对另一个数据集的指导意义非常有限。而基于核函数变换的聚类算法则通过引入核函数将输入空间线性不可分的数据映射到更高维的核空间中使其可分，典型的算法如核 K 均值（KernelK-means）聚类[59]。不过，由于该算法实际上只是在核空间实现了 K 均值聚类，所以原算法所具有的一些缺点仍然被继承了下来。

通过前述分析，每种聚类算法都有其优势和局限。例如，基于划分的方法速度快，但却不能发现不规则的簇结构，而谱聚类虽然能避免该问题，却由于较大的计算量而常被用于局部空间的聚类分析等。同时，在处理一些特殊数据分析时，对聚类分析算法也常有特殊的要求，例如：

（1）聚类分析算法要有发现任意簇形状的能力；

（2）由于数据的分布情形未知，应尽量避免人为设定簇数量；

（3）应具有较好地处理异常数据和高维数据的能力；

（4）聚类结果应具有一定的可解释性、指示价值等。

要解决上述问题，需要从以样本学习为主要内容的统计学习理论中寻找方法，支持向量机作为统计学习理论、VC 维理论和结构风险最小化原则下的优秀算法，它的提出正好迎刃而解了这些问题：首先，它能使用有限的样本来描述所分析的数据对象，而这些样本既可作为簇原型也可作为一部分样本的代表，对进一步的数据分析提供丰富的知识和指示价值；其次，支持向量机也是核方法的一种，它通过核函数将高维特征空间的运算转化为原输入空间的运算，避免了需要显式使用特征空间的数据而产生维数灾难和计算困难的问题。同时，高维特征空间的超平面则对应了输入空间的非线性分界面，有效地将输入空间的非线性问题转化为了特征空间中的线性问题。然后，通过对分界面表达模型的调整，使其能以最小的样本子集来划分任意形状的簇轮廓，这种对数据分布结构的准确定位，对于新的知识发现提供了帮助。最后，它还可通过引入松弛因子来控制风险，并得到全局最优的结果。

由于早期支持向量机为解决分类问题而设计，其目的是追求具有最大分类间隔的分类超平面（见 2.3 节）。1999 年，DavidM.J. Tax 等[60]对其改造后允许

生成具有闭合边界的超平面，使得支持向量机能顺利对单类问题的数据进行描述。之后，Ben-Hur A 等[61,62]进一步引入簇标定算法，提出了支持向量聚类算法原型，其基本步骤如下：

（1）SVC 训练寻找支持函数。通过一个核函数将所有样本点从数据空间映射到特征空间后求解一个凸优化问题，在特征空间中寻找一个能够包围所有数据的最小超球体（设半径为 R），将该球体逆映射回原输入空间后，就会形成表示数据集中分布区边界的轮廓，而轮廓上的数据点就是支持向量（Support Vectors，SV）。在特征空间中，这些 SV 距离超球体中心距离为 R，它们一起构造支持函数。

（2）簇标定（Cluster Labeling）。通过对任意样本对进行采样，若采样点在特征空间中距离超球体中心距离大于 R，则说明该对样本不属于同一个簇，反之则同簇。由于支持向量聚类算法继承了支持向量机的优点，迎合了数据挖掘对聚类的特殊要求，而被成功用于手写体识别[39]、文本聚类[63]、图像分割[48,64]、消除噪点和图像修复[65]、流式数据分析[66]、电力系统负载分析[67]、集成电路制造[68,69]、医学领域（脑部活跃性检查）[70,71]、未知雷达信号检测[72]及关系数据挖掘[73]等领域。

1.4 支持向量机与文本分类

文本分类（Text Categorization，TC）的基本任务是预定义类别或标签集，然后根据文本内容判定它的类别或标签。例如，在数据防泄露保护（Data Leakage Protection，DLP）系统中，由于大部分用户数据以文本形式存在，DLP 系统在定位敏感信息前，需要先定义不同敏感级别（作为类标）的敏感数据样本，通过有监督学习建立具有增量学习能力的分类模型，然后用于新文本信息内容的敏感度判别。因此，一般的文本分类系统包括文本预处理、文本表示和分类器分类 3 步流程，如图 1-1 所示。

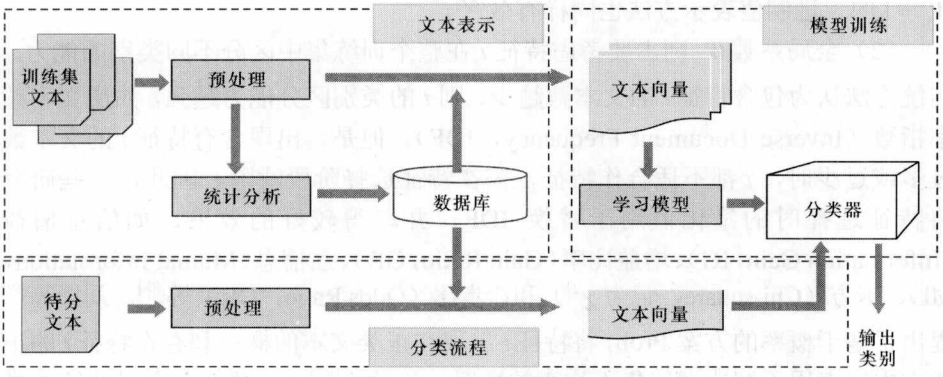


图 1-1 文本分类模型

为强调文本分类系统的一般性，文本预处理工作主要包括去除文档标记、去停用词（Stop Words）、分词或取词根、词性标注、统计词频及数据清洗等。其中，对于欧洲语系的文本，因其可通过空格简易分词，所以一般需要通过词根提取来提供更好的相似度判别（当然，可直接使用近义词、反义词的语义标注来衡量[74]），而对于亚洲语系，则需要专用的分词算法进行处理，如用于中文分词的 ICTCLAS[75]。

文本表示的根本目的是通过向量空间模型（Vector Space Model，VSM）为分类器提供分类对象的量化信息，因此相关研究主要集中于以什么语义单元作为项及计算项的权重问题上[76]。目前，大部分工作采用词或 n-gram 作为项，并以项的频率为基础计算权重。Strzalkowski[77]认为这种简单地词语堆砌模式（Bag of Words，BOW）忽略了词语与词语之间的关系，而只有词按某种规则排列才能明确表示文本的语义内容。之后，陆续有研究人员提出通过本体[78]、句法规则[79]、专业术语、词汇序列（如后缀树）[80]、语义计算[81]及多词组合[82]等增强语言单元的方式改进文本表示的项，然而这些规则的引入，通常和文本分类的具体应用领域紧密相关，并不具备通用性，例如文献[83]在为情感分类服务时建立了大量的情感词、字符串等。最近研究表明，由于具有较强的泛化能力，BOW 模式实际上仍然是目前最主流的文本项模式[18, 84]，于是对文本表示的研究则可集中在如何量化文本特征的权重问题上。通常，特征权重可划分为局部系数 W_{L}、全局系数 W_{G} 和规范化系数 W_{N} 3 个组成部分 [85, 86]：

（1）局部系数 W_{L} 用于表征特征 t 对当前文本 D 的直接影响，并且认为特征 t 在文本 D 中出现的频率（Term Frequency，TF）f 越高，其重要性越大。但是，由于特征的 f 值变化范围差异较大，可直接影响权重范围并关系到欧式空间中的文本相似度计算，因此衍生出一些新的表达式，比如 $\log f$、$\log(f+1)$ 和 $1+\log f$ 等。文献[84]提出通过衡量特征词在文本中分布的均衡程度反映其重要性，结合特征频率与文本长度比值的方式表达局部系数。另外，当 $f>1$ 时，令 $W_{\mathrm{L}}=1$ 的二进制值表示方法也同样有效[87]。

（2）全局系数 W_{G} 则主要考虑特征 t 在整个训练集中区分不同类别的能力。传统方法认为包含特征 t 的文本数越少，则 t 的类别区分能力越强，如逆文本频率指数（Inverse Document Frequency，IDF）。但是，出现含有特征 t 的文本数过多或过少时，t 都不适合作特征，需在特征选择阶段滤除。因此，一些研究将特征选择时的量化值用于替换 IDF，并取得较好的效果，如信息增益（Information Gain，IG）、增益比率（Gain Ratio，GR）、互信息（Mutual Information，MI）、卡方（Chi-square，记为 χ^2）和让步比（Odds Ratio，OR）等[88]。刘影等[89]提出了基于概率的方案 Prob，将特征 t 出现在正类文本的概率和存在特征 t 的正类文本比率用于解决训练集不均衡的情况。兰曼等[87]认为要全局衡量特征的类

别区分能力，只需考虑包含该特征的正类文本数和负类文本数之间的关系，并定义了相关频率（Relevance Frequency，RF）量化该关系用于构建文本向量，该方案被 Altçay 等[90]通过理论分析证明是目前最好的全局系数方案之一。

（3）规范化系数 W_N 的作用是将用 $W_L \times W_G$ 表达的特征权重映射到[0,1]区间，以使得不同长度的文本向量之间具有可比性[86]。通常，使用规范化系数对提升文本分类效果的作用是显著的。

文本分类过程是有监督的学习过程，因此分类器的选择与改进也非常重要。目前，主要的分类器包括 Rocchio[91,92]、决策树（Decision Tree，DT）[3]、朴素贝叶斯分类器（Naive Bayes）[93,94]、神经网络（Neural Networks）[95,96]、K 近邻（K-Nearest Neighbor，KNN）[97]、质心分类器（Centroid Classifier）[98]和支持向量机（Support Vector Machines，SVM）[99,100]等。结合文本分类自身的特点，文献[18,19,84,87]对众多分类器分析和比较后发现，支持向量机和 KNN 分类器更适合于文本分类，尤其以支持向量机分类器精度高，且只需存储支持向量样本用于新文本的分类。相对于其他分类器，支持向量机在文本分类方面的优势主要体现在：

- 支持向量机基于 VC 维理论和结构风险最小化原则，能通过输入空间的运算有效地解决文本数据的高维、稀疏问题；
- 支持向量机对应于二次规划问题，避免了神经网络分类器无法克服的局部最优值问题；
- 文本向量特征之间有明显的相关性，而诸如朴素贝叶斯等建立在特征独立性假设基础上的算法受该特性影响较大，而支持向量机则对此不敏感；
- 支持向量机针对有限样本情况得到最优解，并可有力地支持增量学习和主动学习模式，解决文本分类样本收集困难、内容更新快等问题，更符合生产实际。

因此，支持向量机分类器具有较好的分类性能和泛化能力，并被广泛用于分类领域[101~106]。特别地，台湾大学林智仁等[107~109]基于二次软间隔支持向量机（L2 Soft-Margin Support Vector Machine，L2-SVM）等理论设计的 Liblinear 分类器高效地解决了传统支持向量机分类器的分类速度问题，进一步加速了支持向量机在文本分类领域的应用。不过，由于基于 VC 维理论，支持向量集包含了分类所必需的信息，对支持向量集的分类等价于对全部训练样本的分类[110]，而 Liblinear 分类器仍然提取了大量的训练样本作为支持向量，以至于所选支持向量样本的代表意义和指示价值不足，因此如何进一步减少支持向量的数量来提高分类的速度，同时在避免精度丢失的前提下使得支持向量样本有更强的指代意义以提供对诸如主题模型、文本摘要等领域的应用提供帮助，都具有重要价值。另外，高维稀疏的向量模型是文本分类的特殊问题，它不仅带来了高维数据空间的噪声影响支持向量机的计算，也是造成基于矩阵运算的各种分类器时间消耗和存储空间消耗的根源，因此，如何将高维稀疏向量模型变得低维稠密，也值得进一步研究。