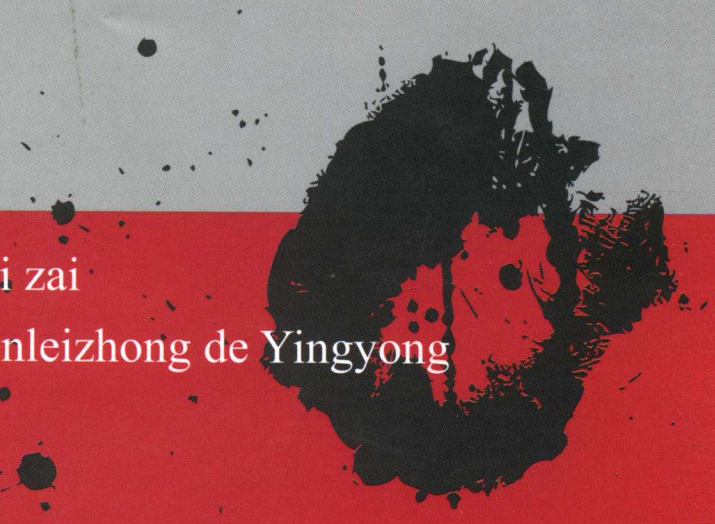
支持向量机
在自动文本分类中的应用

刘 爽 编著



Zhichi Xiangliangji zai
Zidong Wenben Fenleizhong de Yingyong

大连海事大学出版社

支持向量机在自动文本分类中的应用

刘　爽　编著

大连海事大学出版社

图书在版编目（CIP）数据

支持向量机在自动文本分类中的应用 / 刘爽编著. —大连：大连海事大学出版社，2014.7

ISBN 978-7-5632-3031-0

I. ①支… II. ①刘… III. ①文字处理—研究 IV. ①TP391.1

中国版本图书馆 CIP 数据核字（2014）第 152785 号

大连海事大学出版社出版

地址：大连市凌海路 1 号　邮编：116026　电话：0411-84728394　传真：0411-84727996

http://www.dmupress.com　E-mail:cbs@dmupress.com

大连住友彩色印刷有限公司印装　大连海事大学出版社发行

2014 年 7 月第 1 版　2014 年 7 月第 1 次印刷

幅面尺寸：170 mm×240 mm　印张：7.5

字数：131 千字　印数：1～500 册

出版人：徐华东

责任编辑：姜建军　张　华　版式设计：晓　江

封面设计：王　艳　责任校对：杨　淼

ISBN 978-7-5632-3031-0　定价：21.00 元

前 言

随着计算机技术、通信技术、网络技术的飞速发展和 Internet 应用的日益普及，电子文档的数量与日俱增。面对如此浩瀚的信息，人们迫切需要寻找到能够快速、准确获得所需信息的途径。而自动文本分类作为信息过滤、信息检索、搜索引擎、文本数据库、数字化图书馆等领域的技术基础，是目前信息检索与机器学习领域的研究热点之一，有着广泛的应用前景。

支持向量机作为一种基于统计学习理论的新型机器学习方法，较好地解决了非线性、高维数、局部极小点等实际问题，广泛应用于多个研究领域。由于文本分类中文本向量稀疏性大、维数高、特征之间具有较大的相关性，而支持向量机对于特征相关性和稀疏性不敏感，处理高维数问题具有较大的优势，因此，支持向量机非常适用于文本分类问题，在文本分类中具有很大的应用潜力。但是，基于机器学习的自动文本分类是一个非常复杂的信息处理任务，无论在理论上还是在实践上，目前仍然面临着许多亟待解决的难题。

为此，本书将对支持向量机算法及其在自动文本分类中的应用研究做全面、深入的探讨。

本书是在作者博士论文的基础上完成的，全书共分为 6 章。第 1 章主要介绍了支持向量机的发展历史，讨论了支持向量机的基本算法、主要研究内容和各种变形的支持向量机算法，以及支持向量机算法在自动文本分类中的应用研究现状。第 2 章提出了基于遗传算法自动选择参数的加权支持向量机算法，补偿类别差异造成的不利影响，并在模型训练过程中引入遗传算法自动选择惩罚因子和核函数宽度两个参数提高模型的推广性能。第 3 章对自动文本分类相关技术做了详细的回顾和总结，包括自动文本分类流程、文本特征表示及选择、各种特征约简方法比较、文本分类性能评价等。第 4 章在分析文本分类问题统计特性的基础上，利

用支持向量机的分类原理对自动文本分类的机器学习本质进行深入探讨，从文档统计矩阵的角度出发，提出基于最大间隔自动文本分类模型，从理论上证明最大间隔文本分类模型的间隔下界，进而估计最小超球半径和错误率的上界，通过标准测试数据集的实验结果验证模型的有效性。第 5 章提出了进行超文本文档分类的加权直推 SVM 算法。结合文档内容的相似度、超链接结构相似度，提出一种度量一个未标记超文本文档与一类标记超文本文档之间相似性的方法，由此计算每个未标记样本的加权因子，对不同的测试样本赋予不同的惩罚项。区别对待测试样本，细化分类决策面的调整过程，从而得到更为可靠的分类规则。第 6 章对支持向量机算法及其在自动文本分类中的应用进行了总结，并对机器学习未来发展、自然语言处理成果应用等下一步工作做了展望。

面对浩如烟海的互联网文本数据，研究机器学习算法在自动文本分类的应用不仅具有较高的理论意义，而且具有巨大的潜在实际工程应用价值。希望本书的工作能为自动文本分类高效实现打下一定的基础。由于机器学习、数据挖掘技术迅速发展，加之作者水平有限，书中难免有不足之处，敬请广大读者批评指正。

编著者

2014 年 5 月

目 录

第1章　绪论

1.1　课题研究背景和意义

随着信息技术的发展，交通行业逐渐形成了以数字交通为标志的现代交通信息技术格局。水路运输的迅速发展促进了航海科技的发展，数字海洋概念的提出，推进了全球海事资料的数字化、网络化、智能化和可视化，因此以电子形式存在的数据总量急剧增长。如 Lloyd's 国际海事数据库的 Seasearcher.com 提供了 117 000 多艘船的最新信息、4 000 多个港口的相关资料以及 16 300 个公司的联系资料，系统的日更新记录达到 12 000 条。

考察 Internet 上的信息量，其网页数量是考察 Internet 上信息量的一项重要指标，据不完全统计，当前 Internet 上的网页数量已经达到 400 亿张左右。目前仅 Google 搜索引擎所引的网页就高达 80 亿张。根据 John Roth 提出的新摩尔定律，Internet 上的信息量正以每 6 个月翻一番的速度爆炸性地增长。

面对数量如此之大的信息，在搜索突发事故救援、航行通告等重要信息时，缺乏有效的途径和手段，很容易淹没在无边的“信息海洋”中。因此，现代航海越来越迫切地需要能有效地利用这些数据资源的工具。

为了更好地研究、开发这类工具，首先就要来认识其需要处理的对象，即电子数据。总的来说，这些数据具有很强的异质性、多变性和无序性。然而，根据其结构化程度的不同，仍可以将它们归为以下三类：（1）结构化数据。这类数据能够用统一的结构加以表示，如关系数据库中的表结构，可以用不同的字段来表示一个事物的各种属性，还可以通过关键词来表达各个表之间的多种关系。这类数据特点是具有明确定义的（well-defined）语法和明确的语义。（2）半结构化数据。这类数据既不是完全结构化的也不是完全无结构的，如超文本文档和 XML（eXtensional Markup Language）文档。（3）非结构化数据。这类数据是完全无结构的，如自由文本文档、音频流数据、视频流数据等。

从信息的存在形态来看，现实世界中，人们可以获取的大部分信息是存储在半结构化或非结构化的文本文档中的，IBM 与 SearchCafe 公司的调查报告就指出：在现有

的电子信息中文本类信息就占了80%以上。将大量的文本文档组织在一起就构成了所谓的文本数据库，文本数据库中的文本文档可以来自各种数据源，如新闻报道、研究论文、书籍、数字图书馆、电子邮件消息和Web页面，等等。当前，在Internet上已经出现了为数众多的文本数据库，比如，人们熟知的万维网就是一个包含大量网页的、动态的、互连的文本数据库。

对于结构化的数据，人们可以用结构化的查询语言查询得到所需要的数据和信息，例如，在关系数据库中，人们可以使用SQL结构化查询语言来查询所需的数据。然而，文本数据库中的数据要么是半结构化的，要么是非结构化的，以关系代数等为理论基础的传统数据库理论不再适用于这类数据，因此，如何有效地利用文本数据库中的数据成为人们面临的一个难题。

当前，人们主要利用各种搜索引擎提供的关键词匹配功能从文本数据库中获取信息，例如，通过Google搜索引擎，普通的网络用户可以从万维网上寻找自己感兴趣的信息，IEEE 数字图书馆也为科研人员提供了专用的搜索引擎。遗憾的是，这些搜索引擎所使用的关键词匹配功能尚不能完全满足人们的需要。举个例子，假设我们希望从万维网上找到与“文本分类”相关的英文科技类网页，当我们把关键词“text classification”提交给Google搜索引擎后，它竟然返回2 040 000个网页！大致浏览一下，就会发现，在返回的这么多网页中，真正与“文本分类”相关的科技类网页只占很少一部分，而其他的网页都是我们不感兴趣的。

搜索引擎所面临的困难的实质在于搜索引擎对要检索的信息仅仅采用机械的关键词匹配来实现，这种处理过于简单，缺乏必要的知识处理能力和理解能力。因此把信息检索从目前基于关键词的层面提高到基于知识（或概念）的层面，是解决问题的根本和关键。为此，人们提出了文本挖掘的概念（也称为文本数据挖掘或文本数据库知识发现）。一般说来，文本挖掘就是指采用机器学习、模式识别、统计分析等方法从文本文档中提取人们感兴趣的模式或知识的一个非平凡（non-trivial）的过程；这里的过程通常是指一个多阶段的过程，涉及数据准备、模式搜索、知识评价以及反复的修改求精等，而该过程要求是非平凡的，意思是指要有一定程度的智能性、自动性。文本挖掘还可以看作是对传统的数据挖掘或知识发现的一种扩展，然而两者是有区别的，最明显的区别在于前者以半结构化或非结构化的文本文档数据为中心，而后者以结构化数据为中心。当前，文本挖掘已经成为一个重要的研究课题，甚至可以说，如果该课题得以圆满解决，那么Internet上信息资源利用的问题就解决了一大半。

文本挖掘任务包括多个方面，如文本分类、文本聚类、文档重要性和相关性排列、文档集合模式与趋势分析等，其中，文本分类既是文本挖掘的一项重要任务，也是其中的一个基本课题。

一般说来，实现文本分类可以采用人工方式和自动方式。传统的文本分类工作是由特定领域的专家人工完成的，这种人工分类方法需要耗费大量的时间和精力，特别是随着文本信息量的增加，人工方法已经很难满足分类的需求，因此利用计算机对文本进行自动分类的自动文本分类日益受到重视。

文本分类的发展与机器学习的发展密切相关，机器学习新方法的提出对文本分类的发展有直接的推动作用。纵观文本分类发展的历史，每一次机器学习技术的突破都会推动文本分类的发展。支持向量机（Support Vector Machines, SVM）是 20 世纪 90 年代中期在统计学习理论基础上发展起来的一种新型机器学习方法，采用结构风险最小化准则（Structural Risk Minimization, SRM）训练学习机器，建立在严格的理论基础之上，较好地解决了非线性、高维数、局部极小点等问题，成为继神经网络研究之后机器学习领域新的研究热点。而文本分类问题中文本向量稀疏性大、维数高、特征之间具有较大的相关性，支持向量机对于特征相关性和稀疏性不敏感，处理高维数问题具有较大优势，因此，支持向量机非常适用于文本分类问题。Dortmund 大学的 Joachims 将支持向量机用于文本分类，取得了比其他文本分类算法更好的分类效果。并且，支持向量机的分类决策函数只和包含重要分类信息的支持向量有关，是强有力的增量学习和主动学习工具，在文本分类中具有很大的应用潜力，能够解决文本分类中许多其他方法难以解决的问题。

在实际应用中，像文本分类这样类别和样本数目多、噪声多的分类问题非常普遍，针对支持向量机方法在文本分类等实际应用中存在的问题进行深入研究，对于支持向量机、机器学习、文本分类以及与它们相关领域的发展都具有重要意义。

1.2　支持向量机概述

支持向量机简称 SVM，是统计学习理论中最年轻的内容，也是最实用的部分，其核心内容是 V. Vapnik 在 1992 年到 1995 年间提出的，目前仍处在不断发展阶段。SVM 实质上是统计学习理论在实际应用中的一种实现方法，在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，而且可以推广应用到函数估计等其他机器学

习问题中。

1.2.1 支持向量机的发展历史和研究现状

作为SVM的奠基者V. Vapnik早在20世纪60年代就开始了统计学习理论的研究。

1971年，V. Vapnik和A. Chervonekis在“The Necessary and Sufficient Conditions for the Uniforms Convergence of Averages to Expected Value”一文中提出了SVM的一个重要理论基础——VC维理论。

1982年，在*The Estimation of Dependences Based on Empirical Data*一书中，V. Vapnik进一步提出了具有划时代意义的结构风险最小化原理，成为SVM算法的基石。

1992年，Boser、Guyon和V. Vapnik在*A Training Algorithm for Optimal Margin Classifiers*一书中，提出了最优边界分类器。

1993年，Cores和V. Vapnik在*The Soft Margin Classifier*一书中进一步探讨了非线性最优边界的分类问题。

1995年，V. Vapnik在*The Nature of Statistical Learning Theory*一书中，完整地提出了SVM分类。

1997年，V. Vapnik、S. Gokowich和A.Smola发表的“Support Vector Method for Function Approximation, Regression Estimation and Signal Processing”一文中，详细介绍了基于SVM方法的回归算法和信号处理方法。

由于SVM算法的潜在应用价值，吸引了国内外众多学者的注意，如1998年A. Smola在他的博士论文中详细研究了SVM算法中各种核的机理和应用，B. Schlkopf对分类和回归问题提出了*v*-SVM算法，Suykens提出了最小二乘支持向量机，Mangasarian等人提出了广义支持向量机。

国内众多学者也对支持向量机的推广和发展做出了许多贡献，如张学工在参考文献[18]中介绍了统计学习理论与支持向量机，并于2000年翻译出版了V. Vapnik的*The Nature of Statistical Learning Theory*；许建华、张学工2004年翻译出版了V. Vapnik的*Statistical Learning Theory*；邓乃扬、田英杰出版了《数据挖掘中的新方法——支持向量机》;李国正、王猛等翻译了Cristianini等的*An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*；张铃研究了支持向量机理论与神经网络规划算法的关系；卢增祥、李衍达提出了交互支持向量机学习算法；田盛丰、黄厚宽研究了回归型支持向量机简化算法；张文生、王珏等提出了在支持向量机中引入后验概率

的方法以及基于邻域原理计算海量数据的支持向量机算法；孙剑、郑南宁等提出了一种改进的 SMO 算法；周水生、周利华等提出了训练支持向量机的低维 Newton 算法；李建民、张钹等提出了序贯最小优化的改进算法；郭崇慧、孙建涛等提出了广义支持向量机优化问题的极大熵方法；宋杰、唐焕文将支持向量机用于同源寡聚蛋白质分类，等等。国内外学者的积极研究极大地推进了支持向量机的快速发展，目前，支持向量机方法应用于模式分类、时间序列预测、工业控制等诸多领域。

1.2.2　支持向量机基本算法

支持向量机方法的主要优点有：

（1）它是专门针对有限样本情况的，其目标是得到现有信息下的最优解，而不仅仅是样本数趋于无穷大时的最优值。

（2）算法最终将转化成为一个二次优化问题，从理论上说，得到的将是全局最优解，解决了神经网络方法中无法避免的局部极值问题。

（3）算法将实际问题通过非线性变换转换到高维的特征空间，在高维空间中构造线性判别函数来实现原空间中的非线性判别函数，同时它巧妙地解决了维数问题，其算法复杂度与样本维数无关。

支持向量机是从数据分类问题的研究中发展而来的，在数据分类问题中，如果采用通常的神经网络方法，可以简单地描述为：系统随机产生一个超平面并移动它，直到数据集中属于不同类的点正好位于超平面的不同侧面。这种处理机制决定了采用神经网络进行数据分类最终获得的分类超平面将相当靠近训练集中的点，在绝大多数情况下，它并不是最优解。而 SVM 考虑寻找一个满足分类要求的超平面，并使训练集中的点距离分类面尽可能地远，也就是寻找一个分类面使它两侧的空白区域（间隔）最大。

假定大小为 n 的训练样本集 $\boldsymbol{x}_i, y_i$, $i=1, 2, \ldots, n$ 由两类别组成，如果 $\boldsymbol{x}_i \in R^d$ 属于第 1 类，则标记为正（$y_i = +1$）；如果属于第 2 类，则标记为负（$y_i = -1$）。学习的目的是构造一个决策函数，将测试数据尽可能正确地分类。针对训练样本集为线性和非线性两种情况分别讨论。

1.2.2.1　线性情况

如果存在分类超平面 $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$ 使得

$$\begin{cases} \boldsymbol{w}\cdot\boldsymbol{x}_i+b\geq+1, y_i=+1 \\ \boldsymbol{w}\cdot\boldsymbol{x}_i+b\leq-1, y_i=-1 \end{cases}, i=1,2,\cdots,n \tag{1-1}$$

则称训练集是线性可分的，否则称训练集是线性不可分的。

支持向量机是从线性可分情况下的最优分类面发展而来的。其基本思想可用图 1-1 所示的两维情况说明。

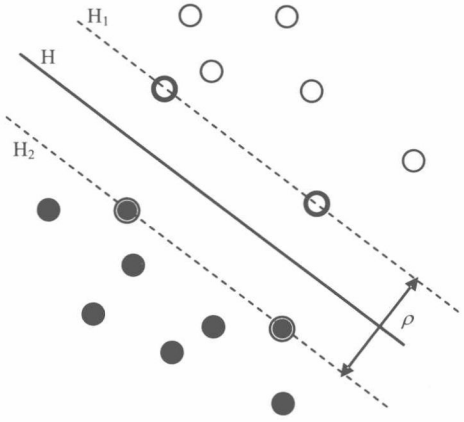


图 1-1　最优分类面示意图

图中，实心点和空心点代表两类样本，H 为分类线，H_1、H_2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫作分类间隔。所谓最优分类线就是要求分类线不但能将两类正确分开，而且使分类间隔最大。

分类线方程为 $\boldsymbol{x}\cdot\boldsymbol{w}+b=0$，对它进行归一化，使得对线性可分的样本集 $\boldsymbol{x}_i, y_i$, $i=1,\ldots,n, \boldsymbol{x}\in R^d, y\in\{+1,-1\}$，满足

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1\geq 0,\ i=1,\cdots,n \tag{1-2}$$

任一点 $\boldsymbol{x}_i$ 到超平面$(\boldsymbol{w}, b)$的距离为

$$d(\boldsymbol{w},b;\boldsymbol{x}_i)=\frac{|\boldsymbol{w}\cdot\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|} \tag{1-3}$$

式（1-2）等价于 $\min_i|(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b|=1$，则由式（1-3）可知

$$d(\boldsymbol{w},b;\boldsymbol{x}_i)\geq\frac{1}{\|\boldsymbol{w}\|} \tag{1-4}$$

分类间隔 ρ 的表达式为：

$$\begin{aligned}\rho(\boldsymbol{w},b) &= \min_{\boldsymbol{x}_i:y_i=-1} d(\boldsymbol{w},b;\boldsymbol{x}_i) + \min_{\boldsymbol{x}_i:y_i=+1} d(\boldsymbol{w},b;\boldsymbol{x}_i) \\ &= \min_{\boldsymbol{x}_i:y_i=-1} \frac{|\boldsymbol{w}\cdot\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|} + \min_{\boldsymbol{x}_i:y_i=+1} \frac{|\boldsymbol{w}\cdot\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|} \\ &= \frac{1}{\|\boldsymbol{w}\|}[\min_{\boldsymbol{x}_i:y_i=-1} |\boldsymbol{w}\cdot\boldsymbol{x}_i+b| + \min_{\boldsymbol{x}_i:y_i=+1} (\boldsymbol{w}\cdot\boldsymbol{x}_i+b)] \\ &= \frac{2}{\|\boldsymbol{w}\|}\end{aligned} \tag{1-5}$$

即此时分类间隔等于 $2/\|\boldsymbol{w}\|$，使间隔最大等价于$\|\boldsymbol{w}\|^2$ 最小，满足式（1-2）且使 $\|\boldsymbol{w}\|^2/2$ 最小的分类面就是最优分类面，H_1、H_2 上的训练样本点称作支持向量。

使分类间隔最大实际就是对推广能力的控制，这是 SVM 的核心思想之一。统计学习理论指出，在 N 维空间中，设样本分布在一个半径为 R 的超球范围内，则满足条件$\|\boldsymbol{w}\| \le A$ 的正则超平面构成的指示函数集 $f(\boldsymbol{x}, \boldsymbol{w}, b) = \text{sgn}(\boldsymbol{w}\cdot\boldsymbol{x} + b)$的 VC 维满足下面的界

$$h \le \min([R^2A^2], N) + 1 \tag{1-6}$$

因此使$\|\boldsymbol{w}\|^2$ 最小就是使 VC 维的上界最小，从而实现 SRM 准则中对函数复杂性的选择。

采用拉格朗日乘子法求解这个具有线性约束的二次规划问题，即

$$\begin{aligned}&\min \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t.}\quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b) \ge 1 \\ &\quad i = 1,\cdots,l\end{aligned} \tag{1-7}$$

对应的拉格朗日多项式为

$$L_{\text{P}} = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\boldsymbol{w}\cdot x_i+b)-1] \tag{1-8}$$

其中，$\alpha_i \ge 0$ 为拉格朗日乘子。由此得到

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0 \tag{1-9}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (1\text{-}10)$$

把式（1-9）、（1-10）带入式（1-8），可以得到上述最优分类面问题的对偶优化问题，即在约束条件

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \qquad (1\text{-}11)$$

$$\alpha_i \geq 0,\ i = 1, \cdots, n \qquad (1\text{-}12)$$

下对 α_i 求解下列函数的最大值

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \qquad (1\text{-}13)$$

α_i 为与每个样本对应的拉格朗日乘子。这是一个不等式约束下二次函数寻优的问题，存在唯一解。容易证明，解中将只有一部分 α_i 不为零，对应的向量就是支持向量。解上述问题后得到的最优分类函数是

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w}^* \cdot \boldsymbol{x}_i + b^*) = \operatorname{sgn}[\sum_{i=1}^{n} \alpha_i^* y_i (\boldsymbol{x}_i \cdot \boldsymbol{x} + b^*)] \qquad (1\text{-}14)$$

由于非支持向量的系数为 0，式中求和实际上只对支持向量进行。

b^* 是分类阈值，可以由任意一个支持向量用式（1-2）求得。为了计算可靠，也可以对所有支持向量分别计算 b 的值，然后求其平均。线性可分的例子见图 1-2。其中带圆圈的点代表支持向量，位于分类超平面的间隔上。

在线性不可分情况下，可以在式（1-2）中增加一个松弛项 $\xi_i \geq 0$，成为

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i \geq 0 \qquad i = 1, \cdots, n \qquad (1\text{-}15)$$

将目标改为求 $(\boldsymbol{w}, \xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{i=1}^{n} \xi_i)$ 最小，即折中考虑最少错分样本和最大分类间隔，就得到广义最优分类面。其中 $C > 0$ 是一个常数，它控制对错分样本惩罚的程度。

图 1-2　线性可分例子图

此时，二次规划问题变为：

$$L_{\mathrm{P}}=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i \tag{1-16}$$

其中，$\beta_i\geq 0$ 为拉格朗日乘子。对 L_{P} 求偏导时除了式（1-9）、（1-10）外，增加一个对 ξ_i 的偏导

$$\frac{\partial L}{\partial \xi_i}=C-\alpha_i-\beta_i=0 \tag{1-17}$$

由此得到广义最优分类面的对偶问题与线性可分情况下几乎完全相同，只是条件（1-12）变为

$$0\leq\alpha_i\leq C,\quad i=1,\cdots,n \tag{1-18}$$

最优化求解得到的 α_i 中，α_i 可能是①$\alpha_i=0$；②$0<\alpha_i<C$；③$\alpha_i=C$。后两者对应的 $\boldsymbol{x}_i$ 为支持向量。由式（1-9）可知只有支持向量对 $\boldsymbol{w}$ 有贡献，也就是对最优超平面、决策函数有贡献，支持向量由此得名，对应的学习方法称之为支持向量机。在支持向量中，②所对应的 $\boldsymbol{x}_i$ 称为标准支持向量（Normal Support Vector, NSV）；③所对应的 $\boldsymbol{x}_i$ 称为边界支持向量（Bounded Support Vector, BSV），实际上是错分的训练样本点。根据 KKT（Karush-Kuhn-Tucker）条件，在最优点拉格朗日乘子与约束的积为 0，即

$$\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]=0$$
$$\beta_i\xi_i=0 \tag{1-19}$$

对于标准支持向量（$0<\alpha_i<C$），由式（1-17）得到$\beta_i>0$，则由式（1-19）得到$\xi_i=0$，因此，对任一标准支持向量，满足

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)=1 \tag{1-20}$$

从而计算参数 b 为

$$b=y_i-\boldsymbol{w}\cdot\boldsymbol{x}_i=y_i-\sum_{x_j\in SV}\alpha_j y_j(\boldsymbol{x}_j,\boldsymbol{x}_i),\quad \boldsymbol{x}_i\in NSV \tag{1-21}$$

其中，SV 为支持向量的集合，NSV 为标准支持向量的集合。为了计算可靠，也可以对所有支持向量分别计算 b 的值，然后求其平均，即

$$b=\frac{1}{N_{\mathrm{NSV}}}\sum_{\boldsymbol{x}_i\in NSV}[y_i-\sum_{\boldsymbol{x}_j\in SV}\alpha_j y_j(\boldsymbol{x}_i,\boldsymbol{x}_j)] \tag{1-22}$$

其中，N_{NSV} 为标准支持向量数。

线性不可分的例子见图 1-3。

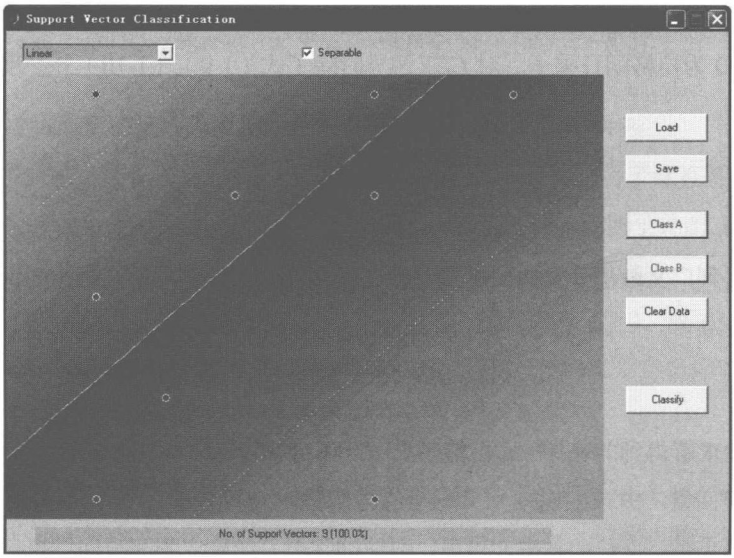


图 1-3　线性不可分例子图

由图 1-3 可见，用线性核函数无法把这两类数据完全分开，明显存在错分现象，此例中存在两个错分样本点，每类错分一个。