中国计算机学会学术著作丛书

——知识科学系列 5

机器学习及其应用2007

周志华　王珏　主编

清华大学出版社

中国计算机学会学术著作丛书
——知识科学系列 5

机器学习及其应用2007

周志华　王珏　主编

清华大学出版社
北京

内 容 简 介

机器学习是人工智能的一个核心研究领域，也是近年来计算机科学中最活跃的研究分支之一。目前，机器学习技术不仅在计算机科学的众多领域中大显身手，还成为一些交叉学科的重要支撑技术。本书邀请相关领域的专家撰文，以综述的形式介绍机器学习中一些领域的研究进展。全书共分 13 章，内容涉及高维数据降维、特征选择、支持向量机、聚类、强化学习、半监督学习、复杂网络、异构数据、商空间、距离度量以及机器学习在自然语言处理中的应用等。

本书可供计算机、自动化及相关专业的研究人员、教师、研究生和工程技术人员参考。

图书在版编目(CIP)数据

机器学习及其应用 2007/周志华，王珏主编. —北京：清华大学出版社，2007.10
（中国计算机学会学术著作丛书. 知识科学系列）
ISBN 978-7-302-16076-2

Ⅰ. 机…　Ⅱ. ①周…　②王…　Ⅲ. 机器学习　Ⅳ. TP181

中国版本图书馆 CIP 数据核字(2007)第 139115 号

责任编辑：薛　慧
责任校对：焦丽丽
责任印制：王秀菊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
c-service@tup.tsinghua.edu.cn
社 总 机：010-62770175　　**邮购热线**：010-62786544
投稿咨询：010-62772015　　**客户服务**：010-62776969
印 刷 者：清华大学印刷厂
装 订 者：三河市李旗庄少明装订厂
经　　销：全国新华书店
开　　本：185×230　**印　张**：18.25　**字　数**：375 千字
版　　次：2007 年 10 月第 1 版　　**印　　次**：2007 年 10 月第 1 次印刷
印　　数：1~3500
定　　价：37.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770177 转 3103　　产品编号：025368-01

评审委员会

名誉主任委员：张效祥
主 任 委 员：唐泽圣
副主任委员：陆汝钤
委　　　员：（以姓氏笔画为序）

王　珊　吕　建　李晓明
林惠民　罗军舟　郑纬民
施伯乐　焦金生　谭铁牛

序

第一台电子计算机诞生于20世纪40年代。到目前为止,计算机的发展已远远超出了其创始者的想象。计算机的处理能力越来越强,应用面越来越广,应用领域也从单纯的科学计算渗透到社会生活的方方面面:从工业、国防、医疗、教育、娱乐直至人们的日常生活,计算机的影响可谓无处不在。

计算机之所以能取得上述地位并成为全球最具活力的产业,原因在于其高速的计算能力、庞大的存储能力以及友好灵活的用户界面。而这些新技术及其应用有赖于研究人员多年不懈的努力。学术研究是应用研究的基础,也是技术发展的动力。

自1992年起,清华大学出版社与广西科学技术出版社为促进我国计算机科学技术与产业的发展,推动计算机科技著作的出版,设立了"计算机学术著作出版基金",并将资助出版的著作列为中国计算机学会的学术著作丛书。时至今日,本套丛书已出版学术专著近50种,产生了很好的社会影响,有的专著具有很高的学术水平,有的则奠定了一类学术研究的基础。中国计算机学会一直将学术著作的出版作为学会的一项主要工作。本届理事会将秉承这一传统,继续大力支持本套丛书的出版,鼓励科技工作者写出更多的优秀学术著作,多出好书,多出精品,为提高我国的知识创新和技术创新能力,促进计算机科学技术的发展和进步作出更大的贡献。

中国计算机学会

2002年6月14日

前　言

机器学习致力于“利用经验来改善系统自身的性能”。在计算机系统中，“经验”通常是以数据的形式存在的，要利用经验就不可避免地要对数据进行分析。因此，机器学习已逐渐成为计算机数据分析技术的源泉之一。随着人类收集和存储数据能力的不断增长以及计算机运算能力的飞速发展，利用计算机来分析数据的要求越来越广泛，越来越迫切，从而使得机器学习的重要性越来越显著。2001 年，美国航空航天局 JPL 实验室的科学家在 *Science* 上撰文指出，机器学习对科学研究的整个过程正起到越来越大的支持作用，并预计该领域将取得稳定而快速的发展；2003 年，美国国防部高级研究计划局(DARPA)开始启动了以机器学习为核心的 PAL 计划，将机器学习技术的重要性上升到国家安全的高度来考虑；2006 年，美国卡内基梅隆大学专门成立了机器学习系。这些情况表明，机器学习已经成为计算机科学技术中最受关注的领域之一。

2002 年，陆汝钤院士在复旦大学智能信息处理实验室发起并组织了“智能信息处理系列研讨会”，并将“机器学习及其应用”列为当年支持的研讨会之一。2002 年 11 月，研讨会成功举行，并确定了会议不征文、不收费，以及“学术至上，其他从简”的办会宗旨。2004 年 11 月，在复旦大学举行了第二届“机器学习及其应用”研讨会，两天半的会议一直有 100 余人旁听，这令与会专家深受鼓舞，于是商定从此次会议开始，将“机器学习及其应用”发展成为一个系列研讨会，在每年 11 月上旬的一个周末举行。2005 年 11 月，南京大学计算机软件新技术国家重点实验室举办了第三届研讨会，吸引了来自全国近十个省市的 250 余人旁听；2006 年 11 月，南京大学计算机软件新技术国家重点实验室和南京航空航天大学信息科学与技术学院联合举办了第四届研讨会，吸引了来自全国十余个省市的 300 余人旁听。同时，为了促进研究生之间以及研究生与资深学者之间的交流，在第四届研讨会期间还举行了“第一届机器学习及其应用学生研讨会”，由一些受到邀请的研究生介绍自己的研究成果，以夜间墙展的方式进行，也吸引了 100 余人参加。

清华大学出版社对推介信息科学技术领域的研究进展一直抱有极大的热情。早在“第二届机器学习及其应用研讨会”举行期间清华大学出版社就参与其中，并为该研讨会专门出版了文集，即 2006 年出版发行的《机器学习及其应用》一书。2005 年第三届研讨会期间，清华大学出版社和与会专家商定，以后每两届研讨会的部分内容将编辑成书，以

《机器学习及其应用：出版年》的形式冠名。

本书是清华大学出版社邀请第三届和第四届“机器学习及其应用研讨会”的部分专家将其报告内容总结成文所得的文集。书中各章按作者的姓氏拼音为序，每一章将讨论一个论题，以综述的形式对该方面的研究进展加以介绍，并将报告人自己的一些研究工作嵌入其中。书中章节不仅涉及支持向量机、聚类分析、特征选择、维数削减、强化学习等传统研究领域，还涉及到流形学习、半监督学习、异构数据分析、商空间等新领域，以及图像理解、网络分析、自然语言处理等应用问题。需要注意的是，书中各章的内容仅表达该章作者本人的见解，并不代表清华大学出版社、编者及其他各章作者的学术观点。本书的出版得到了陆汝钤院士的支持和指导，并得到清华大学出版社计算机专著出版基金的资助，在此谨表示衷心的感谢。

编　者

2007 年 3 月

目　录

1 图像空间中的距离*

封举富　王立威
北京大学信息科学技术学院，视觉与听觉信息处理国家重点实验室，北京　100871

1.1　引言

给定两幅图像，如何度量它们的差异或相似性，这是图像识别和检索的基本问题。一幅 $M\times N$ 大小的数字图像通常以矩阵的方式来表示和存储：

$$\boldsymbol{X}=\begin{bmatrix} x^1 & \cdots & x^N \\ \vdots & \ddots & \vdots \\ x^{(M-1)\times N+1} & \cdots & x^{M\times N} \end{bmatrix} \tag{1}$$

一般要把二维图像扩展为 $\mathbf{R}^{M\times N}$ 空间中的一个点或者向量 $\boldsymbol{x}=(x^1,\cdots,x^{M\times N})$。这样，一幅图像对应 $\mathbf{R}^{M\times N}$ 空间中的一个点，我们把图像所对应的空间称为图像空间。一组图像或者一段视频图像在 $\mathbf{R}^{M\times N}$ 空间张成一个子空间或者流形。如在人脸图像识别过程中，在各种姿势和光照等条件下，可以得到大量的人脸图像，它们生成了图像空间中的一个低维人脸子空间[1]。在视频识别中，数据的基本单位是一段视频，也就是由很多帧组成的一系列图像。衡量两段视频的相似程度就要求给出两组图像间的距离度量。

在图像空间中可以建立各种距离。严格意义上的距离度量满足非负性、自反性、对称性和三角不等式等性质。本文主要介绍图像空间中两幅图像间的距离和两组图像间的距离。

1.2　两幅图像间的距离

度量图像距离的方法有很多[2,6,9,10,14~20]，最常用的是欧氏距离。图像空间中，两幅图像的欧氏距离就是两个点之间的欧氏距离：

---

*　本文得到国家自然科学基金(60575002,60635030)、国家重点基础研究发展计划(2004CB318000)和新世纪优秀人才支持计划资助。

$$d_{ED}(x,y)=\left(\sum_{i=1}^{M\times N}(x^i-y^i)^2\right)^{1/2} \tag{2}$$

更一般的，有 Minkowski 距离：

$$d_{MD}(x,y)=\left(\sum_{i=1}^{M\times N}|x^i-y^i|^p\right)^{1/p} \tag{3}$$

通常也称为 L_p 范数。L_1 范数，也称为 Manhattan 距离或城区距离。L_∞ 范数就是 $|x^i-y^i|$ 的最大值。

二值图像可以看做二维图像平面上的点集，因此，可以借鉴数学上关于点集之间距离的定义。给定欧氏空间中的两个点集 $A=\{a_1,\cdots,a_m\}$，$B=\{b_1,\cdots,b_n\}$，可以定义它们之间的 Hausdorff 距离：

$$\begin{aligned}H(A,B)&=\max[h(A,B),h(B,A)]\\ h(A,B)&=\max_{a\in A}\min_{b\in B}\|a-b\|\end{aligned} \tag{4}$$

$h(A,B)$通常被称作有向 Hausdorff 距离，它给出了点集 A 中的点到点集 B 的最大距离。$h(B,A)$则给出了点集 B 中的点到点集 A 的最大距离。因此，Hausdorff 距离度量了点集 A 与 B 之间的最大差异。当 A，B 都是闭集时，Hausdorff 距离满足非负性、自反性、对称性和三角不等式。当 A，B 都是凸集时，Hausdorff 距离等于它们的边缘之间的 Hausdorff 距离。因此，在用 Hausdorff 距离来度量图像距离时，一般要先进行边缘提取。但是，Hausdorff 距离对于噪声非常敏感。当图像存在噪声时，它们将成为图像差异最大的点。由此，Huttenlocher 等人[2]提出了广义 Hausdorff 距离来度量二值图像的距离。广义 Hausdorff 距离采用第 k 个最大值来决定距离。其最大的特点是允许图像在有遮挡的情况下（即部分图像）进行比较。该距离现已广泛应用于二值图像以及形状的比较[3~5]。

人的视觉对于图像的几种典型变化，如图像的平移、旋转、伸缩等具有很强的鲁棒性。当图像发生上述某种变化时，所有这些由变化生成的图像在图像空间中构成的集合是什么样的呢？如果变化只取决于一个参数（比如横向平移的参数，就是平移的距离），那么图像变化的轨迹就是一条曲线。如果变化有 k 个自由度（比如二维空间中的刚体运动有 3 个自由度），那么所有经变化得到的图像构成图像空间中的 k 维流形。必须要注意的是，即使是平移和旋转这样简单的变换，得到的图像流形也是非线性的。为了使图像距离对于变换具有不变性，一个很自然的想法是定义图像距离为两个流形之间的距离。但是，由于流形是非线性的，求解流形间的距离非常困难。另一方面，有时候我们只希望图像距离并非对于变换总是不变的，只具有局部不变性——也就是对微小的变化具有不变性就够了。比如“6”和“9”两个数字，如果图像距离对旋转完全不变，那么这两个数字的距离是零，这并不是我们希望的结果，因此局部不变显得更加合理。

Simard 等人[6]正是基于上述考虑在研究手写数字识别时提出了切距离（tangent distance）。

它首先考虑两幅图像 P 与 E 的变化流形在两幅图像处的切空间。显然，如果流形是 k 维的，则切空间也是 k 维的。切距离就定义为两个切空间之间的最小距离。由于两个线性子空间之间的距离就是最小二乘问题，因此非常容易解出。在手写数字应用中，切距离考虑 7 种变换：纵横两个方向的平移、旋转、伸缩、线条粗细以及两个双曲变换（保证 6 种二维平面仿射变换都包括在内）。采用切距离的最近邻方法在著名的美国邮政服务手写数字数据集（USPS）上，达到了与人的识别率相同的水平，大大超过其他方法。该方法还被推广到一般的模式识别问题[7,8]上。但是，计算切距离需要充分利用先验知识，针对具体图像考虑可能的变换，不同类型的图像一般具有不同的变换。尽管切距离在手写数字识别中取得了很大的成功，但其应用还是受到一定的限制。

图像的先验知识不仅可以表示为确定的信息，也可以由模糊集合的隐身概念引出。模糊图像度量 FIM（fuzzy image metric）[9]就是这方面的代表。模糊图像度量基于模糊积分的概念，改进像传统欧氏距离那样逐像素对比的方法，考虑一个像素及其周围某个邻域内所有像素的关系，并且应用于图像压缩与编码，取得了较好的效果。

我们继续考虑图像空间中的欧氏距离。当图像由矩阵形式扩展为向量形式时，图像所具有的几何空间关系消失了，因此欧氏距离对图像噪声和变化非常敏感，即使图像仅有微小变化，也可能产生很大的距离数值。一个很自然的想法就是把图像的几何空间关系纳入到距离度量中。

首先，以 $e_1,\cdots,e_{M\times N}$ 作为图像空间的一组基，其中 e_{kN+l} 对应于位于 (k,l) 的一个理想点信号源。于是一幅图像 $\boldsymbol{x}=(x^1,\cdots,x^{M\times N})$，其中 x^{kN+l} 是 (k,l) 像素的灰度值，就对应于图像空间中的一个点，并且 x^{kN+l} 是 e_{kN+l} 的坐标值。图像空间的原点代表一幅灰度值处处为零的图像。

图像空间的距离度量完全由基向量的度量系数确定。度量系数 $g_{ij}, i,j=1,\cdots,M\times N$ 定义为

$$g_{ij} = \langle \boldsymbol{e}_i,\boldsymbol{e}_j\rangle = \sqrt{\langle \boldsymbol{e}_i,\boldsymbol{e}_i\rangle}\sqrt{\langle \boldsymbol{e}_j,\boldsymbol{e}_j\rangle}\cdot\cos\theta_{ij} \tag{5}$$

其中，$\langle,\rangle$ 代表标量积，θ_{ij} 是基向量 $\boldsymbol{e}_i$ 和 $\boldsymbol{e}_j$ 之间的夹角。注意，如果 $\langle \boldsymbol{e}_i,\boldsymbol{e}_i\rangle=\langle \boldsymbol{e}_j,\boldsymbol{e}_j\rangle=\cdots$，即所有基向量的长度都相等，则 g_{ij} 完全依赖于夹角 θ_{ij}。给定了度量系数，两幅图像的距离度量就可以写成

$$d(x,y) = \left(\frac{1}{2\pi}\sum_{i,j=1}^{M\times N} g_{ij}(x^i-y^i)(x^j-y^j)\right)^{1/2} = (x-y)^{\mathrm{T}}\boldsymbol{G}(x-y) \tag{6}$$

其中对称矩阵 $\boldsymbol{G}=(g_{ij})_{MN\times MN}$ 称为度量矩阵。

由此，我们提出了图像欧氏距离[10]。

定义 1　一个距离度量 $d(x,y)=[(x-y)^{\mathrm{T}}G(x-y)]^{1/2}$，$\boldsymbol{G}=(g_{ij})_{MN\times MN}$ 称为图像欧氏距离 IMED（image euclidean distance），如果它满足条件：

(C1) 度量系数 g_{ij} 完全由像素 P_iP_j 间距离决定。我们用 f 来表示这一函数关系：

$$g_{ij} = f(|P_i - P_j|), \quad i,j = 1,\cdots,M\times N \tag{7}$$

(C2) f 是连续函数，而且 g_{ij} 随 $|P_i - P_j|$ 的增加而单调减小；

(C3) f 是普适函数，也就是说，它适用于任意大小和分辨率的图像。

C1 要求像素距离信息必须包含于图像距离度量中。g_{ij} 仅由 $|P_i - P_j|$ 决定使得距离对于图像的线性变换具有不变性。同时，这还意味着所有的基向量具有相同长度，因此，g_{ij} 与 $\cos\theta_{ij}$ 具有线性关系。C2 指出了如何将像素距离信息融合到度量系数中，从而使诱导的图像距离直观上是合理的。f 的连续性是一个一般性的要求。g_{ij} 随 $|P_i - P_j|$ 的增加而单调减小意味着图像距离与形变的程度有关。最后，C3 保证了定义的度量对所有图像都适用。

为了使条件 C1～C3 全部得到满足，f 必为正定函数。最重要的正定函数是 Gauss 函数。

$$g_{ij} = f(|P_i - P_j|) = \frac{1}{2\pi\sigma^2}\exp(-|P_i - P_j|^2/2\sigma^2) \tag{8}$$

其中 σ 是方差参数，在本文中总是设为 1。由式(2)和式(8)可以看出，当 $\sigma\to\infty$ 时，图像欧氏距离趋于传统的欧氏距离。实际上，当二维图像扩展为高维空间中的点或向量时，传统的欧氏距离没有考虑图像像素之间的空间关系，其基向量是完全独立的。而图像欧氏距离考虑了图像像素之间的空间关系，并体现在度量系数中。

我们利用它来构造图像欧氏距离(IMED)：

$$d_{\mathrm{IMED}}(x,y) = \left(\frac{1}{2\pi}\sum_{i,j=1}^{M\times N}\exp(-|P_i - P_j|^2/2)(x^i - y^i)(x^j - y^j)\right)^{1/2} \tag{9}$$

进一步考虑矩阵 $\boldsymbol{G}$ 的分解：$\boldsymbol{G}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$。尽管存在着无穷多种形如 $\boldsymbol{G}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 的分解，我们采用其中十分特殊的一种：

$$\boldsymbol{G} = \boldsymbol{G}^{1/2}\boldsymbol{G}^{1/2} \tag{10}$$

其中，对称矩阵 $\boldsymbol{G}^{1/2}$ 唯一地被下式定义：

$$\boldsymbol{G}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^{\mathrm{T}} \tag{11}$$

这里，$\boldsymbol{\Lambda}$ 是以 $\boldsymbol{G}$ 的特征根为元素的对角阵(由于 $\boldsymbol{G}$ 正定，因此 $\boldsymbol{\Lambda}^{1/2}$ 的对角元素都是正实数)，$\boldsymbol{\Gamma}$ 是正交阵，其列向量是 $\boldsymbol{G}$ 的特征向量。由此，图像欧氏距离诱导了一个变换，我们称之为标准化变换(standardizing transform)：

$$u = \boldsymbol{G}^{1/2}x, v = \boldsymbol{G}^{1/2}y \tag{12}$$

图像欧氏距离实质上也是变换域上的欧氏距离：

$$d_{\mathrm{IMED}}(x,y) = ((x-y)^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(x-y))^{1/2} = ((u-v)^{\mathrm{T}}(u-v))^{1/2} \tag{13}$$

我们目前的大部分识别方法都是基于欧氏距离的。因此，所有这些识别方法都可以移植到这个变换域中来。也就是说，可以先对图像做一个标准化变换，然后利用这些识别