支持向量机的若干新进展

王国胜,钟义信
(北京邮电大学信息工程学院,北京 100876)

摘 要: 支持向量机是九十年代中期发展起来的机器学习技术,与传统的人工神经网络不同,前者基于结构风险最小化原理,后者基于经验风险最小化原理.实验表明,支持向量机不仅结构简单,而且技术性能尤其是泛化能力明显提高.本文是一篇综述,介绍支持向量机研究的一些新进展,希望引起大家的重视.

关键词: 支持向量机; 模式识别; 算法
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2001) 10-1397-04

Some New Developments on Support Vector Machine

WANG Guo-sheng, ZHONG Yi-xin
(School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Support vector machine is new machine learning technique developped from the middle of 1990s. Being different from traditional neural network, it is based on structure risk minimization principle, while the latter on empirical risk minimization principle. A large number of experiments have shown that, comparing with traditional neural network, support vector machine has not only simpler structure, but also better performances, especially better generalization ability. In this paper, some new developments on support vector machine are introduced so as to draw our attention.

Key words: support vector machine; pattern recognition; algorithm

1 引言

支持向量机(Support Vector Machine,简称 SVM)是 Cortes&Vapnik1995 年首先提出来的[4],是近年来机器学习研究的一项重大成果.根据 Vapnik&Chervonenkis 的统计学习理论[12~14],如果数据服从某个(固定但未知的)分布,要使机器的实际输出与理想输出之间的偏差尽可能小,则机器应当遵循结构风险最小化原理,而不是经验风险最小化原理,通俗地说就是应当使错误概率的上界最小化.支持向量机正是这一理论的具体实现.与传统的人工神经网络相比,支持向量机不仅结构简单,而且各种技术性能尤其是泛化(generalization)能力明显提高,这已被大量实验证实[2,8].目前,国内对支持向量机的研究还刚刚起步,本文介绍支持向量机的一些新进展,希望引起大家的重视.

2 支持向量机

什么是支持向量机?我们从模式识别问题谈起.

设输入模式集$\{x_i\} \subset R^n$由两类点组成,如果x_i属于第一类,则标记 1,如果属于第二类,则标记 - 1.从中取大小为ℓ的样本作为训练集(x_i, y_i), $i = 1, 2, \cdots, \ell$,这里$y_i = 1$或 - 1.学习的目标是要构造一个判别函数,将两类模式尽可能正确地区分开来.分三种情况讨论.

2.1 线性可分情形

定义 1 称训练集是线性可分的,如果存在分离超平面$(w \cdot x) + b = 0$,使得

$$\begin{aligned}(w \cdot x_i) + b \geq 1, \ y_i = 1\\ (w \cdot x_i) + b \leq -1, \ y_i = -1\end{aligned} \tag{1}$$

可把以上不等式合并写成

$$y_i[(w \cdot x_i) + b] \geq 1, \ i = 1, 2, \cdots, \ell \tag{2}$$

注 对一个固定的超平面,参数(w, b)不是唯一确定的(相差一个常数因子).因此总能够找到一对(w, b),使上述不等式中至少有一个以等式成立,为此只须令$\{(x_i, y_i)\}$到该超平面的最小距离 = $1/\|w\|$.这样的表示方式称为这个超平面关于$\{(x_i, y_i)\}$的标准表示.我们约定,以下的超平面均采用标准表示.

定义 2 称分离超平面是最优的,若训练集到该超平面的最小距离最大.

根据定义及上面的注,$(w \cdot x) + b = 0$是最优的当且仅当(w, b)是下面优化问题的解

$$\begin{aligned}&\min \quad \frac{1}{2}\|w\|^2\\ &\text{s.t.} \quad y_i[(w \cdot x_i) + b] \geq 1, \ i = 1, 2, \cdots, \ell\end{aligned} \tag{3}$$

这个二次规划问题有唯一的极小点,用 Lagrange 乘子法把(3)

收稿日期:2000-04-03;修回日期:2001-01-21
基金项目:国家自然科学基金(No. 69982001)

化成其对偶形式

$$\max \quad \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j}^{\ell}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \tag{4}$$

$$\text{s.t.} \quad \sum y_i\alpha_i = 0,\ \alpha_i \geq 0,\ i = 1,2,\cdots,\ell$$

和

$$w = \sum_{i=1}^{\ell} y_i\alpha_i x_i$$

若 $\alpha_i > 0$,称相应的 x_i 为支持向量. 于是最优超平面方程为 $\sum_{x_i \in sv}\alpha_i y_i (x_i \cdot x) + b = 0$. 采用判别函数

$$y = \operatorname{sgn}[\sum_{x_i \in sv}\alpha_i y_i (x_i \cdot x) + b] \tag{5}$$

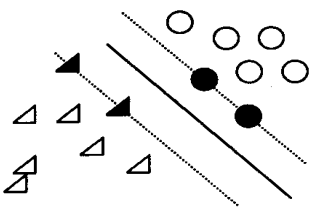

图 1 实直线为最优超平面

一般来说,在实际问题中支持向量仅占训练样本的一小部分(如图 1 中虚直线上实心圆和三角),这样(5)中被加项就较少了.

2.2 线性不可分情形

若训练集是线性不可分的,或事先不知道它是否线性可分,引入非负变量 ξ_i, $i = 1,2,\cdots,\ell$,与式(3)相对应的优化问题是

$$\min \quad \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{\ell}\xi_i$$

$$\text{s.t.} \quad y_i[(w \cdot x_i) + b] \geq 1 - \xi_i \tag{6}$$

$$\xi_i \geq 0,\ i = 1,2,\cdots,\ell$$

这里 ξ_i 可看作训练样本关于(广义)分离超平面的偏差,$\xi_i = 0$ 时问题变为线性可分情形,$c > 0$ 是自定义的惩罚系数,用来控制样本偏差与机器泛化能力(与 $\frac{1}{2}\|w\|^2$ 有关)之间的平衡.

用 Lagrange 乘子法把式(6)化成其对偶形式

$$\max \quad \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \tag{7}$$

$$\text{s.t.} \quad \sum_{i=1}^{\ell} y_i\alpha_i = 0, 0 \leq \alpha_i \leq c,\ i = 1,2,\cdots,\ell$$

和

$$w = \sum_{i=1}^{\ell} y_i\alpha_i x_i \tag{8}$$

若 $\alpha_i > 0$,称相应的 x_i 为支持向量. 判别函数为

$$y = \operatorname{sgn}[\sum_{x_i \in sv}\alpha_i y_i (x_i \cdot x) + b] \tag{9}$$

2.3 支持向量机

超平面的分类能力毕竟有限,为此引入分离曲面,主要思想是:作非线性映射 $\Phi(x): R^n \to F$, F 是高维内积空间称为特征空间,$\Phi(x)$ 称为特征映射;然后在 F 中构造(广义)最优超平面.

推导过程与 2.2 完全相同,只是把那里的 x, x_i 分别换成 $\Phi(x)$, $\Phi(x_i)$. 不难看出为了求分离曲面,不必知道 $\Phi(x)$ 的确切表达式,只要知道如何由输入 x, z 计算内积 $(\Phi(x) \cdot \Phi(z))$ 就够了,即

$$(\Phi(x) \cdot \Phi(z)) = k(x,z) \tag{10}$$

这个事实对构造支持向量机有重要意义,当特征空间的维数巨大时(实际情况常常如此,有时甚至是无穷维的),若直接计算内积其复杂度可想而知,因此行不通. 上述事实指出:高维特征空间中内积运算,可转化为低维输入空间(相对而言维数要低得多)上一个简单的函数计算.

称对称函数 $k(x,z)$ 为核函数,称判别函数

$$y = \operatorname{sgn}[\sum_{x_i \in sv}\alpha_i y_i k(x_i \cdot x) + b] \tag{11}$$

为支持向量机(如图 2 所示).

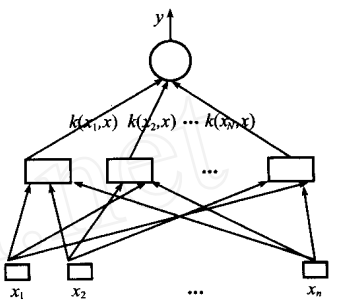

图 2

支持向量机由训练集和核函数完全刻画,这样在实际问题中,常常是直接给出核函数(而不是先给出映射 $\Phi(x)$),所以如何构造、选择核函数是个重要问题. 给定输入空间 R^n 上二元函数 $k(x,z)$,怎么判断它是否核函数(即是否存在某个内积空间 F 和映射 $R^n \to F$,使(10)式成立)? 著名的 Mercer 定理给出一个充分条件[12]. 常用的核函数有

多项式核 $k(x,z) = ((x \cdot z) + t)^p$, $t \geq 0$, p 是自然数.

高斯(径向基)核 $k(x,z) = e^{-(\|x-z\|^2/2\sigma^2)}$, $\sigma > 0$.

两层神经网络核 $k(x,z) = S(a(x \cdot z) + t)$,其中 S 是 sigmoid 函数,a, t 是某些常数.

实验表明,支持向量机有以下特点:

® 结构简单.

® 性能优良,尤其是泛化能力好.

® 适合处理高维数据:计算复杂性与输入模式的维数没有直接关系,避免了维数灾难.

® 有关的优化问题有唯一的极小点.

® 学习速度快.

® 更换核 k,可以得到各种不同的分离曲面.

支持向量机不仅用于模式识别,这里的方法已被推广到回归估计,函数逼近等[11,15].

3 简化的支持向量机

在关于手写体数字识别的比较研究中[9],人们发现虽然支持向量机在成功率方面有明显优势,但识别速度不及某些神经网络. 为解决这个问题,Burge 提出了一种简化的支持向

量机[3],不久这一技术被 Scholkopf 等进一步发展[9]. 鉴于该问题在基于核的回归分析,主成分分析中同样存在,下面针对一般情形进行描述.

设在特征空间中

$$\Psi = \sum_{i=1}^{\ell} \alpha_i \Phi(x_i) \tag{12}$$

这里 x_i 是输入模式或支持向量,(回想 2.3 及式(7), $w = \sum_{i=1}^{\ell} \alpha_i y_i \Phi(x_i)$ 中系数 $\alpha_i y_i$ 相当于这里的 α_i). 为了减少计算量,试图寻找 Ψ 的一个近似 $\Psi' = \sum_{i=1}^{m} \beta_i \Phi(z_i)$, $m < \ell$,使得 $\| \Psi - \Psi' \|^2$ 最小.

注意这里 z_i 不必是支持向量. 提出这个问题,不仅出于主观上的要求,而且有现实可能性. 第一,如果训练集中有若干相同样本,因为展开系数必须满足某种约束(如 $0 \le \alpha_i \le c$,见优化问题(7)),因此式(12)中可能出现同类项,将其合并就能得到简化的展开式. 第二,实际问题中训练集是各种各样的.

上述问题涉及两方面,一是确定最优系数 β,二是确定简化集 $\{z_i\}$.

定理 1[9]　设 $\Phi(z_i)$ 线性无关, $K_{ij}^z := (\Phi(z_i) \cdot \Phi(z_j))$, $K_{ij}^{zx} := (\Phi(z_i) \cdot \Phi(x_i))$,则最优系数 $\beta = (K^z)^{-1} K^{zx} \alpha$.

注　倘若 $\Phi(z_i)$ 线性相关,则总可以从中去掉若干个,使余下的向量是线性无关的. 不过这里的 β 依赖于 z_i,设想如果找到一个算法它能够同时优化 β 和 z_i,那么用这些 z_i 和定理 1 所求得的最优系数至少和 β 一样好. 从这个意义上说,定理 1 是有用的.

如何寻找 z_i 呢? 主要有两种方法,第一从 $\{x_i\}$ 中选出 m 个最重要的作为 z_i,这种方法的优点是算法简单,但实际效果不好;第二从 R^n 中选. 实验表明,第二种方法比第一种要好得多,所以下面只介绍第二种方法. 先考虑一种简单情况.

3.1　$\Psi' = \beta \Phi(z)$

这时为了最小化 $\| \Psi - \Psi' \|^2$,只要使 Ψ 到 $\{\Phi(z)\}$ 张成的子空间的距离最小,即最小化

$$\left\| \frac{(\Psi \cdot \Phi(z))}{(\Phi(z) \cdot \Phi(z))} \Phi(z) - \Psi \right\|^2 = \| \Psi \|^2 - \frac{(\Psi \cdot \Phi(z))^2}{(\Phi(z) \cdot \Phi(z))},$$

为此只要最大化 $\frac{(\Psi \cdot \Phi(z))^2}{(\Phi(z) \cdot \Phi(z))} = \frac{(\Psi \cdot \Phi(z))^2}{k(z,z)}$. 这时一旦求出 z,则最优系数 $\beta = \frac{(\Psi \cdot \Phi(z))}{(\Phi(z) \cdot \Phi(z))}$. 假设核有如下形式(如高斯核)

$$k(x,z) = k(\| x - z \|^2), \quad \text{且 } k(z,z) = 1.$$

则问题变为最大化 $(\Psi \cdot \Phi(z))^2$. 令 $\nabla_z (\Psi \cdot \Phi(z))^2 = 2(\Psi \cdot \Phi(z)) \nabla_z (\Psi \cdot \Phi(z)) = 0$ 得到充分条件

$$\nabla_z (\Psi \cdot \Phi(z)) = 0 \tag{13}$$

顺便指出,倘若 $(\Psi \cdot \Phi(z)) = 0$,则 z 不可能是最大值点,除非 $(\Psi \cdot \Phi(z))^2 \equiv 0$,我们对这种退化情形不感兴趣. 将 $\Psi = \sum_{i=1}^{\ell} \alpha_i \Phi(x_i)$ 代入(13)并整理得 $\sum_{i=1}^{\ell} \alpha_i k'(\| x_i - z \|^2)(x_i - z) = 0$, $z = \frac{\sum_{i=1}^{\ell} \alpha_i k'(\| x_i - z \|^2) x_i}{\sum_{i=1}^{\ell} \alpha_i k'(\| x_i - z \|^2)}$. 可通过下面的迭代求出 z

$$z_{n+1} = \frac{\sum_{i=1}^{\ell} \alpha_i k'(\| x_i - z_n \|^2) x_i}{\sum_{i=1}^{\ell} \alpha_i k'(\| x_i - z_n \|^2)}.$$

特别地对高斯核 $k(x,z) = e^{-(\| x_i - z \|^2 / 2\sigma^2)}$,有 $z_{n+1} = \frac{\sum_{i=1}^{\ell} \alpha_i e^{-(\| x_i - z_n \|^2 / 2\sigma^2)} x_i}{\sum_{i=1}^{\ell} e^{-(\| x_i - z_n \|^2 / 2\sigma^2)}}$.

以上算法称为 PA 算法(Preimage Algorithm).

3.2　简化集的构造

算法如下:

(1)设 $\Psi_1 = \sum_{i=1}^{\ell} \alpha_i \Phi(x_i)$,用 PA 算法求 β_1, z_1 使 $\| \Psi_1 - \beta_1 \Phi(z_1) \|^2$ 最小.

(2)令 $\Psi_2 = \Psi_1 - \beta_1 \Phi(z_1)$,用 PA 算法求 β_2, z_2 使 $\| \Psi_2 - \beta_2 \Phi(z_2) \|^2$ 最小. 然后根据定理 1 用 $\Phi(z_1), \Phi(z_2)$ 重新计算最优系数 β_1, β_2(为了叙述简单,优化后的系数不妨仍记为 β_1, β_2).

(3)一般地设 $\Psi_m = \sum_{i=1}^{\ell} \alpha_i \Phi(x_i) - \sum_{i=1}^{m-1} \beta_i \Phi(z_i)$,用 PA 算法求 β_m, z_m 使 $\| \Psi_m - \beta_m \Phi(z_m) \|^2$ 最小. 然后根据定理 1 用 $\Phi(z_1), \Phi(z_2), \cdots, \Phi(z_m)$ 重新计算最优系数 $\beta_1, \beta_2, \cdots, \beta_m$. 当迭代达到预定的次数,或者误差 $\| \Psi_{m+1} \|$(即 $\| \Psi_1 - \sum_{i=1}^{m} \beta_i \Phi(z_i) \|$)小于预定的门限,迭代过程停止. 最后得到简化的展开式 $\Psi' = \sum_{i=1}^{m} \beta_i \Phi(z_i)$ $(m < \ell)$.

从第二步开始,每次都要重新计算最优系数,这是由于一般情况下, $\Phi(z_1), \Phi(z_2), \cdots, \Phi(z_m)$ 不是正交的,因而通过正交投影不能得到 Ψ 的最佳近似.

Scholkopf 等用以上算法,构造了简化的支持向量机,并在 USPS 手写体数字数据库上进行了实验(选用高斯核)[9]. 结果表明,识别速度提高到未简化时的 10 倍,错误率仅从 4.4%上升到 5.1%,综合指标高于传统神经网络. 实验还与通过整体优化 (β_i, z_i) 构造简化集的方法[3]进行了比较,虽然后者的错误率降至 5.0%,但计算量比这里的方法大约高两个数量级.

4　构造与实际问题有关的核

核函数与支持向量机的性能有密切关系,如何构造与实际问题有关的核函数,一直是支持向量机研究的重要课题. Amari &Wu 设计了一种算法[1],他们通过对核函数的黎曼几何分析,提出利用实验数据逐步修正已有的核函数,使之更好地与实际问题相吻合.

设特征映射 $U = \Phi(x)$,则 $dU = \sum_i \frac{\partial}{\partial x_i} \Phi(x)\, dx_i$(注意 $\Phi(x)$ 是高维向量), $\| dU \|^2 = \sum_{i,j} g_{ij}(x)\, dx_i dx_j$,这里 $g_{ij}(x) =$

$(\frac{\partial}{\partial x_i}\Phi(x))\cdot(\frac{\partial}{\partial x_j}\Phi(x))$. 称非负定距阵 $(g_{ij}(x))$ 为 R^n 上(由 $\Phi(x)$ 诱导)的黎曼张量, $ds^2=\sum_{i,j}g_{ij}(x)\,dx_i dx_j$ 为 R^n 上的黎曼距离. 赋予黎曼距离的 R^n 成为黎曼空间, 体积微元 $dV=\sqrt{g(x)}\,dx_1\cdots dx_n$, 其中 $g(x)=\det(g_{ij}(x))$. 通俗地说, $g(x)$ 反映了特征空间中点 $\Phi(x)$ 附近局部区域被放大的程度, 因此也称 $g(x)$ 为放大因子. 因为 $k(x,z)=(\Phi(x)\cdot\Phi(z))$, 可以验证[1] $g_{ij}(x)=\left.\frac{\partial k}{\partial x_i\partial z_j}\right|_{z=x}$. 特别地对高斯核 $k(x,z)=e^{-(\|x-z\|^2/2\sigma^2)}$, $g_{ij}(x)=\frac{1}{\sigma^2}\delta_{ij}$.

在模式识别问题中, 为了更好地将两类不同的模式区分开, 希望尽量拉大他们之间的距离. 确切地说尽量放大分离曲面附近的局部区域, 而保持其他区域变化不大. 但是, 人们事先并不知道分离曲面是什么. 考虑到支持向量几乎总是出现在分离曲面附近, 故设法放大支持向量的局部区域, 可用修正核函数的办法达此目的.

设 $c(x)$ 是 R^n 上正的可微实函数, $k(x,z)$ 是高斯核函数. 根据 Mercer 定理[12]

$$\tilde{k}(x,z)=c(x)\,k(x,z)\,c(z) \tag{14}$$

也是核函数, 可计算出相应的 $\tilde{g}_{ij}(x)=c_i(x)\,c_j(x)+c^2(x)\,g_{ij}$, 这里 $c_i(x)=\frac{\partial}{\partial x_i}c(x)$. Amari &Wu 设 $c(x)$ 有下面的形式[1]

$$c(x)=\sum_{x_i\in SV}h_i e^{-(\|x_i-x\|^2/2\tau^2)} \tag{15}$$

$\tau>0$ 是参数, h_i 是权系数. 则在支持向量 x_i 附近[1] $\sqrt{\tilde{g}(x)}\approx\frac{h_i^n}{\sigma^n}e^{-(nr^2/2\tau^2)}\sqrt{1+\frac{\sigma^2}{\tau^4}r^2}$, 这里 $r=\|x-x_i\|$ 是欧几里得距离. 为保证 $\sqrt{\tilde{g}(x)}$ 在 x_i 附近(即 r 很小)取最大值, 同时在其他区域取较小值, 经计算知

$$\tau\approx\sigma\sqrt{n} \tag{16}$$

于是新的训练过程由两步组成:

(1) 先用某个核 k(高斯核)进行训练, 然后按照式(14), (15)和(16)得到修正的核 $\tilde{k}$.

(2) 用 $\tilde{k}$ 进行训练.

实验表明[1], 这种改进的训练方法不仅可以明显地降低错误识别率(几乎降低了一半), 而且意想不到的是还能减少支持向量的个数, 从而提高了识别速度.

5 结束语

当前, 对支持向量机的研究方兴未艾, 本文介绍了部分有代表性的工作, 这方面的工作还有许多, 总的来说, 主要围绕两个方面: 一是通过对支持向量机本身性质的研究, 提出进一步完善的措施, 如文中所介绍的那样, 此外还包括多类识别问题和快速训练算法等. 二是不断探索新的应用领域, 支持向量机本质上是一种非线性数据处理工具, 人们注意到它在数字信号处理、图象处理、智能控制等领域有巨大的应用潜力, 这方面已经有了一些结果, 如基于核的主成分分析、非线性去噪、非线性模式重建以及数据挖掘等. 支持向量机是一项很有发展前途的技术.

参考文献:

[1] S Amari, S Wu. Improving support vector machine classifier by modifying kernel function [J]. Neural Networks, 1999, 12: 783 - 789.

[2] P L Bartlett, J Shawe-Taylor. Generalization performance on support vector machines and other pattern classifiers [C]. in B. Sholkopf, C. Burges, and A. Smola Eds., Advances in Kernel Methods-Support Vector Learning, Cambridge, MA: MIT Press, 1999.

[3] C J C Burges. Simplified support vector decision rule [C]. in Proc. 13th Int. Conf. Machine Learning, San Mateo, CA, 1996: 71 - 77.

[4] C Cortes, V Vapnik. Support vector networks [J]. Machine Learning, 1995, 20: 273 - 295.

[5] F Girosi, M Jones, T Poggio. Regularization theory and neural networks architectures [J]. Neural Comput., 1995, 7(2): 219 - 269.

[6] Massimiliano Pontil, Alessandro Verri. Properties of support vector machines [J]. Neural Comput., 1998, 10: 955 - 974.

[7] B Sholkopf, A Smola, K R Muller. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Comput., 1998, 10: 1229 - 1319.

[8] B Sholkopf, K Sung, C J C Burges, et al. Comparing support vector machine with Gaussian kernels to radial basis function classifiers [J]. IEEE Trans. Signal Processing, 1997, 45: 2758 - 2765.

[9] B Scholkopf, S Mika, C J C Burges, et al. Input space versus feature space in kernel-based methods [J]. IEEE. Trans. Neural Networks, 1999, 10(5).

[10] B Scholkopf, S Mika, A Smola, et al. kernel PCA pattern reconstruction via approximate preimages [C]. in Proc. 8thInt. Conf. Artificial Neural Networks, Perspectives in Neural Computing, L. Niklasson, M. Boden, and T. Ziemke. Eds. Berlin Germany: Springe-Verlag, 1998: 147 - 152.

[11] A Smola, B Scholkopf. A tutorial on support vector regression [R]. NeuroCOLT, Rep. 19, 1998.

[12] V N Vapnik. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[13] —Statistical learning theory [M]. New York: Wiley, 1998.

[14] —An overview of statistical learning theory [J]. IEEE Trans. Neural Networks, 1999, 10(5).

[15] V N Vapnik, S Golowich, A J Smola. Support vector method for function approximation, regression, and signal processing [C]. in: NIPS 9, San, 1997.

作者简介:

王国胜 男. 1966 年生. 北京邮电大学在读博士生. 主要研究方向有机器学习、神经网络.

钟义信 男. 1940 年生. 北京邮电大学副校长, 博士生导师. 主要研究方向有信息科学理论、人工智能与神经网络.