智能计算研究进展与发展趋势*

黄德双

（中国科学院合肥物质科学研究院智能机械研究所 合肥 230031）

摘要 智能计算技术是一门涉及物理学、数学、生理学、心理学、神经科学、计算机科学和智能技术等的交叉学科，近年来发展迅猛。本文简要介绍了智能计算的学科背景、原理和特点，评述了国际的发展现状和趋势。

关键词 智能计算，研究现状，发展趋势

黄德双研究员

1 引言

智能计算技术是一门涉及物理学、数学、生理学、心理学、神经科学、计算机科学和智能技术等的交叉学科。目前，智能计算技术在神经信息学、生物信息学、化学信息学等交叉学科领域得到了广泛应用。这项技术所取得的些许进步，都会进一步促进神经信息学、生物信息学、化学信息学等交叉学科的发展，反过来，后者的深入研究和进一步发展，也将大大促进智能计算技术的长足进步。所以，深入开展智能计算技术研究具有重要意义，应引起我们的高度关注。

智能计算技术是将问题对象通过特定的数学模型进行描述，使之变成可操作、可编程、可计算和可视化的一门学科。它运用其所具有的并行性、自适应性、自学习性来对信息、神经、生物和化学等学科中的海量数据进行规律挖掘和知识发现。由于其在整个计算过程中自始至终考虑计算的瞬时性和敏捷性，因而对于复杂的问题对象能够通过任务分解或变换方法，使得问题对象能够在有限的时间内获得令人满意的解。

过去，智能计算技术的进步总是离不开人工智能，特别是人工神经网络技术的发展，但是以符号推理为特征的人工智能技术由于过于依赖规则，以至被认为缺少数学支持而遭到质疑；而以自学习、自适应、高度并行性为特征的人工神经网络技术，虽有坚实的数学支撑但又无法精确处理实际问题中的各种小样本集事件，这些大大限制了智能计算技术的进一步发展。近年来，由于支撑向量机（Support Vector Machine：SVM）、核（Kernel）方法和征战模型（Divide-and-Conquer：DAC）等新方法的相继出现，使智能计算技术发展成不但能处理海量数据等大样本集的问题对象，同时也能自适应地处理小样本事件集的数据，从而使该项技术更切合实际需求，更受人们的广泛青睐。

2 国际发展现状及趋势

2.1 智能计算模型

信息技术的发展离不开经典数理统计

* 收稿日期：2005 年 11 月 18 日

学,而智能计算技术的每一进步更以数理统计学为灵魂。众所周知,数理统计学的本质是以 Bayes 理论为基础、对随机事件或过程进行规律统计或挖掘,其中事件或样本的概率密度函数是 Bayes 理论得以广泛应用的基石。由 Bayes 理论为基础,进一步出现了各种提取有用信息或信号的估计方法,如著名的 Weiner 滤波器、Kalman 滤波器等等。一般来说,在以 Bayes 理论为框架的信息处理方法中,最基本的一个假设是,所要解决的问题对象的样本数必须具有一定的规模,以至在信号传输或处理过程中能够估计或近似估计有用信号的概率密度函数,而且理论证明,基于 Bayes 理论的有用信号的无偏估计误差能够达到克拉美·罗 (Cramer–Rao) 的下界。尽管如此,这是一个非常理想的情况,因为很多实际的问题对象很难得到大样本的数据集,如手写签名识别、信用卡防伪验证、人脸识别和语音识别等等。

以人工神经网络为代表的非线性“黑箱”处理模型,尽管对无法用数学模型精确描述的问题的处理具有其独特的优势,但对小样本数据集问题却很难训练网络收敛,且网络求解或描述问题的精度非常低,即使对大样本数据集问题能够使网络训练收敛,但往往会出现过拟合情况,而且有时需要设计非常庞大的网络结构来适应。即便如此,网络收敛后的输出也只是近似 Bayes 后验概率估计。也就是说,在极限情况下,神经网络能够逼近 Bayes 理论的估计。

近年来,以 Vapnik 的支撑向量机为代表的统计学习理论是专门研究小样本情况下的机器学习规律,它采用结构风险最小化准则,在最小化经验风险的同时最小化推广误差的上界,并采用凸二次规划技术使所求问题的解具有全局最优性。SVM 不仅能保证在小样本条件下仍具有较好的推广性,而且基本消除了“维数灾难”、“过学习”和“欠学习”等传统机器学习方法难以解决的问题,在复杂系统的建模、优化、分类与预测等方面显示出强大的优势,使得 SVM 理论成为当前学术界研究的热点问题,受到普遍关注和重视。不过,SVM 在应用中存在两个突出的问题,即 SVM 核函数的选择和 SVM 用于多类问题的学习算法设计。此外,实际应用中很难获得高质量、大规模的数据样本,数据样本中或包含不完整数据、或样本数很少、或蕴含模式多样性的情况。如何充分利用有限数据样本和不完整数据样本中包含的有限信息,构造高精度的 SVM 分类器是一个有待深入研究的问题。

近年来,围绕上述三大模型应用的主要发展趋势是:(1)先验信息的充分利用。即根据所求解问题的先验信息来选择确定具体的统计模型,或者是将问题的先验信息耦合到具体模型中以构造约束模型来求得问题的解。如在神经网络输出误差代价函数中将问题的先验信息通过拉格郎日乘子耦合进来,以构造一种新的约束学习算法来加快问题的求解;(2)任务分解和输出集成。即对于复杂问题,先将整个问题分解成若干个子问题,并由具有较大差异的模型来分别处理,然后通过集成方法把每个子任务对应的模型的输出进行综合,以获得问题的满意解。如在分类器集成研究中,我们拟寻找差异性较大的单个分类器,然后使用 Boosting 算法进行集成,以获得最佳的分类效果。

2.2 特征提取

在实际应用中,我们所得到的数据不但非常庞大,而且非常复杂,有时甚至存在各种冗余,因此在选择具体模型进行处理(如分类或预测)前,有必要首先对这些数据进行一定的分析,如进行一定的变换以提取数据中的主要特征,以利于后面的分析与处

理。

Fisher 线性判别分析（FLDA）是由 Fisher 于 1936 年提出的用于两类问题特征提取的一种有效方法，其基本思想是寻找一投影方向，使训练样本投影到该方向时尽可能具有最大类间距离和最小类内距离。后来，人们又将两类问题的 FLDA 方法推广到多类情况，其基本原理是通过寻找一投影矩阵使得训练样本经投影变换后尽可能具有最大类间散射和最小类内散射。不过，由于 LDA 是线性特征提取方法，因此一般只适用于线性可分的模式。但实际应用中，许多模式并非线性可分，因此，LDA 方法并不理想。为了解决非线性模式的有效特征提取问题，一种可能的办法是对 LDA 方法进行相应的非线性扩展。近年来，随着统计学习理论，特别是支撑向量机(SVM)方法的问世，通过再生核理论对一些线性算法进行非线性扩展已成为研究非线性特征提取方法的一种非常重要的手段。

继 Scholkopf 等人提出了核主成份分析（KPCA）以及 Mika 等人针对两类的 FLDA 问题提出了核 Fisher 判别分析(KFDA)之后，Baudat 等人利用核技巧推广了多类的 LDA 方法，提出了广义判别分析(GDA)方法。目前，GDA 方法已广泛用于指纹、虹膜、人脸等生物特征识别领域，并取得甚至比 SVM 更好的实验结果。此外，同神经网络、SVM 等其它智能计算方法相比，GDA 方法具有计算简单、推广性能好等诸多优点。由于 GDA 本质上是 LDA 在 Hilbert 再生核空间上的扩展，因此 LDA 方法存在的某些本质问题同样会出现在 GDA 中，而且还可能更加突出，其中主要的问题包括奇异性问题、秩限制问题和简并特征值扰动问题。通常解决这些问题的办法是分阶段的方法，亦即通过两种或多种组合技术来解决，如 PCA+LDA，PCA+GDA 等等。

此外，近年来在神经信息学、生物信息学、化学信息学等学科领域还出现了典型相关分析（CCA）、偏最小二乘(PLS)、Logistic 回归等多元统计数据处理技术，而且它们也被推广用来实现判别分析。

随着支持向量机理论的提出，基于核的学习方法已逐渐受到人们的重视。核学习已经远远超越 SVM 范畴，形成了一个相对独立的研究方向，并走向更为广阔的舞台。目前已出现了 Kernel based PCA (KPCA), Kernel based CCA (KCCA)、Kernel based LDA (KLDA)以及 Kernel based Clustering (KC)等特征提取算法。模式分析核方法的中心思想是，在进行分类等数据处理时，对于线性不可分样本，首先通过一个非线性映射将原空间样本映射到高维特征空间（也称核空间）中，使核空间中的样本变得线性可分，或者近似线性可分，然后在核空间中用线性方法对其进行处理，从而实现相对于原空间进行非线性的处理，其效果相当好。

目前，核方法中的核函数主要包括径向基函数（RBF）核、多项式（Polynomial）核和 Sigmoidal 核等。不过，在实际应用中，到底选择什么样的核函数才能最好地变换或表达该问题，还是一个尚未解决的问题。

2.3 模型估计

在实际问题中还经常会遇到来自多个总体并按一定比例混合的数据，这种数据的建模和分析一直是模式识别、聚类分析和信号处理等领域中的一个重要内容，在神经信息学、生物信息学、化学信息学等交叉学科领域有着广泛的应用。对于有限混合体模型参数估计的研究可追溯到 19 世纪末 Pearson 的工作。但从 Pearson 开始到上世纪 60 年代，人们所使用的主要是矩方法和最大似然法等经典方法。这些方法仅仅对一些

特殊混合体分布的参数估计有效。直到1977年,Dempster等建立的期望最大(EM)算法才为一般混合体分布的参数估计提供了一种统一的理论框架。近年来,人们沿着这个方向做了很多努力并建立了许多改进的算法。然而,这些方法的前提是混合体模型中分量个数的选择必须正确,否则将导致错误的参数估计结果。不过,在许多情况下数据的分量个数是未知或难于准确地知道,这时该模型的参数估计就变得异常困难。

在上世纪70年代,Akaike针对有限混合体模型中分量个数的选择问题提出了著名的Akaike信息准则。随后,人们对这一准则进行了多种推广。这种方法是相当耗时的,因为需要对每一个可能的k值进行一次参数估计,并根据这些估计结果计算信息或价值函数以选择最优的k值。这种大量重复计算特别是对于高维大批量数据的情况就更困难。因此,在实际应用过程中人们一直在呼吁自动模型选择方法,也就是通过一次优化过程达到参数估计和模型选择的双重目的,这种方法在速度上将大大优于过去的信息或价值准则方法。该方法将对模式识别、聚类分析和信号处理等领域产生重要的影响,并给实际应用带来方便和快捷。

本质上,有限混合体模型的自动模型选择问题是从观察数据直接推测模型阶数和参数的技术,而目前正在蓬勃发展的独立分量分析(ICA)技术是一种从观察数据的角度探索发射(送)源独立信号个数并分离的技术,它们在图象特征提取、基因微阵数据分析等方面正得到广泛应用。特别是,如果信号传输的信道存在非线性环节,对应的ICA就变成了盲源分离(BSS)技术。目前ICA或BSS发展的"瓶颈"是如何解决高度非线性混合模式的解混,以及如何求解混合矩阵是奇异矩阵、源信号的个数大于观察信号的个数(即overcomplete问题)等问题。

2.4 学习算法

学习算法是对问题解的寻优过程。现实中几乎所有的系统或模型在实际应用前都需要根据输入数据样本来对自身进行学习或训练,以便系统或模型能记住或熟悉所训练的输入模式,然后对未知的样本模式进行测试和评判等。因此,学习算法研究是智能计算技术研究中的一个非常重要的环节。

自1944年Hebb提出改变神经元连接强度的Hebb规则开始,即首次出现了"学习算法"的概念。1957年,Rosenblatt首次引进了感知器(Perceptron)的概念,并正式引进了"学习算法"。1962年,Widrow提出了自适应线性元件(ADLINE),并提出了自适应最小均方(LMS)学习算法。1974年,Werbos在其博士论文中第一次提出了能够实现多层网络训练的反向传播(BP)算法,可以说是"学习算法"史上的一次革命。不过,由于BP算法本质上就是LMS算法,因而其存在局部极小值、训练速度慢等缺陷。随后,出现了大量改进的BP算法,以及一些变型的学习算法等。上世纪80年代初又出现了遗传算法(GA)和模拟退火算法(SAA)等,从而部分解决了局部极小值问题,并大大加快了算法的收敛速度。特别是,GA还能用来解决非数值问题的全局寻优问题,因而推广了学习算法的应用范围。近年来,又出现了一些新的群体学习算法,如黄蜂优化算法(SOA)、免疫算法(IA)、粒子群优化算法(PSOA)以及小生境(niche)技术等等。这些算法都是基于群体的随机搜索技术,实际上也是一种进化计算技术。它起源于对鱼群、鸟群的捕食行为和社会认知模式的模拟。同遗传算法相比,这些算法相对简单和更容易实现,并且没有太多参数需要调整,这些算法近年来得到国内外学者的广泛重视和研究,并获得了一定范围

的应用。现在国际上每年举办一届群智能研讨会,专门讨论群体学习与优化理论及应用方面的研究进展。

3 我国研究进展

我国特别是中科院非常重视智能计算技术的理论与应用研究,并采取措施推动这项技术在我国的发展。2003 年由合肥智能机械研究所、自动化研究所联合清华大学在北京举办了"生物信息学与进化计算"第 81 次青年科学家论坛,吸引了全国 30 多名生物信息学和智能计算领域的青年科学家参会并做专题报告。论坛还专门邀请了清华大学李衍达院士做了大会报告,他介绍了生物信息学与智能计算学科的发展趋势。2005 年由合肥智能机械研究所、中国科技大学,联合香港浸会大学举办了第一届国际智能计算学术会议,会议吸引了 39 个国家和地区的 2 400 多名学者踊跃投稿,专门邀请了美国、英国和香港等著名学者做关于国际上智能计算领域最新发展趋势的大会报告,另外还特别邀请了中科院半导体研究所王守觉院士就智能仿生模式识别问题做了专题演讲。此次大会的成功召开,标志着我国在智能计算相关领域的学术研究已处于国际先进水平。

3.1 模型估计

关于有限混合体模型的混合数自动确定的问题(或称为自动模型选择问题),香港中文大学的徐雷于 1993 年提出一种被称作"对手惩罚竞争学习(RPCL)"算法。RPCL 算法本质上是一种竞争学习算法,可用于数据的聚类分析。它不同于以往的竞争和其它聚类分析方法,能够在估计模型参数的过程中自动确定出数据中的类别个数。随后,徐雷教授还提出了"贝叶斯阴阳学习系统"理论,建立了另一个衡量有限混合体模型建模的和谐函数。通过优化这种和谐函数得到模型参数的估计,同样能够实现有限混合体数据的自动模型选择。但是目前这种方法及其改进形式只在高斯有限混合的情况下才有效,其基本理论问题还远远没有解决,有效的目标代价函数和学习算法还没有寻找到,如何获得一种有效的学习算法来实现参数估计和分量个数的自动确定是自动模型选择问题的一个重要研究方向。

3.2 特征提取

在数据或样本处理领域,近年来,南京理工大学杨静宇教授等在主成份分析(PCA)的基础上提出了一种时间更快、计算效率更高的二维 PCA(简称 2DPCA)。实验结果表明 2DPCA 特征提取效果至少要好于 PCA,不过,2DPCA 要求的内存比 PCA 大。该工作发表在 *IEEE Transaction on Pattern Analysis and Machine Intelligence*(Vol 26, No.1,2004)上。随后,在 2DPCA 的启发下,北京交通大学袁保宗教授等又提出了二维 LDA (简称 2DLDA)。该工作发表在 *P.R.L.* (No.3,2005)上。2DPCA 和 2DLDA 给人的启发是,一些看似很古老的问题仍然可以找到较新的解决途径,此外 2D 技术更加适合图像(或者矩阵)数据的处理,因为它本身是处理二维数据的,因此对于指纹、虹膜、人脸等图像特征提取是有较大意义的。事实上,一维推广到二维的本质是由向量到向量的投影变成矩阵到向量的投影。因此,我们也可能基于这一思想将 CCA、 PLS 等推广到 2DCCA、2DPLS,以及其它更为复杂的情况。

在模式识别领域,中科院半导体所王守觉院士领导的研究小组从另一角度进行了探索。王院士认为,人类是基于对同类事物的共同属性的认识区分不同事物的。近年来他们以"认识"事物而不是"区分"事物为目的研究了模式识别问题,提出了仿生模式识别理论。与传统的以"最佳划分"为目标的统

计模式识别相比，该理论更接近于人类“认识”事物的特性，他们称之为“仿生模式识别”。该理论认为，同类样本在特征空间中的分布具有数学连续性（不能分裂成两个彼此不邻接的部分），即所谓同源连续性原理。采用“仿生模式识别”理论及“高维空间复杂几何形状覆盖神经网络”的识别方法，能得到很高的识别率。

2005 年中科院合肥智能机械研究所黄德双领导的小组，在智能计算方面的研究成果，以封面文章的形式发表在 *Digital Signal Processing: A Review Journal*(Vol.15, No.4, 2005)上。这篇文章主要是讨论使用模糊 c-均值聚类(FCMC)和核主分量分析（KPCA）方法，对该实验室于 2003 年从美国乔治·华盛顿大学带回来的用 U2 飞机拍摄的地面“7 通道多频谱遥感彩色图像数据”进行特征提取和对比度增强处理所取得的重要结果。评审人认为，这是一项漂亮的工作，它把 FCMC 和 KPCA 结合起来，能够很好地实现图像数据中非线性模式变量的特征提取，而且计算量大为减少。特别是，所提取的非线性特征的个数可以多于主分量分析提取的线性特征的个数，从而可以有效提取图像中的弱信息，即分布数量较少的目标信息。另外，通过对比度增强方法，能明显提高非线性特征图像的质量。

3.3 数学机械化

上世纪 70 年代后期，中科院吴文俊院士提出了使用机器帮助人们实现数学定理自动证明的思想，这为数学机械化奠定了坚实的基础。由机器来代替人实现自动化智能化处理，是人类孜孜以求的目标。吴文俊院士在这方面做出许多创新性研究成果，如非线性代数方程组求解的吴方法、偏微分代数方程组的整序方法等等，这些成果已经应用到包括机器人机构的位置分析、智能计算机辅助设计、图像压缩等领域。

实际上我们还可以进一步将“机器证明”推广到更一般的“机器解题”领域。考虑现实中的每个问题总蕴涵一定的先验信息，机器解题中的一个关键问题是，如何使得机器在求解问题时能自动运用来自问题中的先验信息，以帮助机器解题并加快求解速度。事实上我们可以借鉴神经网络中权值的自适应学习办法来解决这一问题。如果将对应的先验信息通过某种形式，耦合到所定义的误差代价函数中，则所推导出来的算法在空间搜索时，必将沿着问题先验信息所指定的方向前进，直至预定的误差。结果所花的搜索时间必然要比未考虑任何先验信息的算法快得多。目前，大量的实验研究已经证实我们的想法。

2004 年中科院合肥智能机械研究所黄德双领导的研究小组在先验信息编码的约束学习算法方面的工作，以封面文章的形式发表在 *Neural Computation* (Vol.16, No.8, 2004)上。这篇文章介绍了用一种新的基于问题先验信息的约束学习算法，来加快神经网络求根器训练速度所取得的重要结果。评审人认为，所提出的约束学习算法是对神经计算领域的重要贡献，它不但对一类求根问题有潜在影响，而且对一类神经计算问题的实时求解也具有重要意义。

4 结语

智能计算技术是信息技术、神经信息学、生物信息学、化学信息等学科发展的核心和基础，它的突破将可能对其它交叉学科产生深远的影响。目前我院在这个领域的研究水平基本处于国内领先地位，已经逐渐成为我国发展智能计算技术最重要、最活跃的研究基地。

主要参考文献

1 Apnik V V. The Nature of Statistical Learning Theory. New York: Springer-Verlag Press, 1995.

2 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine learning, 1997, 29 (2-3): 131-163.
3 Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 1 299-1 319.
4 Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation, parallel distributed processing: Explorations in the Microstructure of cognition, Vol I: Foundation. Cambridge, Massachusetts:MIT Press, 1986.
5 Zhan-Li Sun, D S Huang, Yiu-Ming Cheung. Extracting nonlinear features for multispectral images by FCMC and KPCA. Digital Signal Processing, 2005, 15(4): 331-346.
6 D S Huang, Horace H S Ip, Zheru Chi. A neural root finder of polynomials based on root moments. Neural Computation, 2004, 16(8):1 721-1 762.
7 D S Huang. A constructive approach for finding arbitrary roots of polynomials by neural networks, IEEE Transactions on Neural Networks, 2004, 15 (2): 477-491.
8 Werbos P J. Beyond regression: New tools for predictions and analysis in the behavioral science. Ph D Thesis, Harvard University, 1974.
9 Xu L, Krzyzak A, Oia E. Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. IEEE Trans. on Neural Networks, 1993,(4): 636-648.

The Development and Prospects of Intelligent Computing Technology

De Shuang Huang

(Intelligent Computing Lab, Hefei Institute of Intelligent Machines, CAS, 230031 Hefei)

Intelligent computing technology is a cross discipline field involved in physics, mathematics, physiology, psychology, neural science, computer science and intelligent technology, etc. Recent years, it is being greatly developed at a larger pace. This paper, first of all, briefly introduces the background, principle and features of intelligent computing subjects. Secondly, the state of arts and prospects for intelligent computing technology is overviewed. After that, the progress made in our country for this discipline is concisely surveyed. Finally, some viewpoints and ideas about this discipline in our academy and country are suggested.

Keywords intelligent computing, state of arts, prospect

黄德双 中国科学院合肥物质科学研究院智能机械研究所研究员，中国科学院研究生院教授，中国科技大学博士生导师、兼职教授。2000 年度中科院“百人计划”入选者。在国内外学术刊物与会议上发表论文 240 余篇，其中 *SCI* 收录论文 70 余篇，*SCI* 他引 130 余次，出版专著和论文集 5 部。获第八届全国优秀科技图书奖二等奖 1 项，省部级奖二、三等奖各 1 项。已培养博士 8 名，硕士 5 名；在站博士后 1 名。