模式识别研究进展

刘成林，谭铁牛

中国科学院自动化研究所 模式识别国家重点实验室

北京中关村东路 95 号

摘要

自 20 世纪 60 年代以来，模式识别的理论与方法研究及在工程中的实际应用取得了很大的进展。本文先简要回顾模式识别领域的发展历史和主要方法的演变，然后围绕模式分类这个模式识别的核心问题，就概率密度估计、特征选择和变换、分类器设计几个方面介绍近年来理论和方法研究的主要进展，最后简要分析将来的发展趋势。

1. 前言

模式识别(Pattern Recognition)是对感知信号（图像、视频、声音等）进行分析，对其中的物体对象或行为进行判别和解释的过程。模式识别能力普遍存在于人和动物的认知系统，是人和动物获取外部环境知识，并与环境进行交互的重要基础。我们现在所说的模式识别一般是指用机器实现模式识别过程，是人工智能领域的一个重要分支。早期的模式识别研究是与人工智能和机器学习密不可分的，如 Rosenblatt 的感知机[1]和 Nilsson 的学习机[2]就与这三个领域密切相关。后来，由于人工智能更关心符号信息和知识的推理，而模式识别更关心感知信息的处理，二者逐渐分离形成了不同的研究领域。介于模式识别和人工智能之间的机器学习在 20 世纪 80 年代以前也偏重于符号学习，后来人工神经网络重新受到重视，统计学习逐渐成为主流，与模式识别中的学习问题渐趋重合，重新拉近了模式识别与人工智能的距离。模式识别与机器学习的方法也被广泛用于感知信号以外的数据分析问题（如文本分析、商业数据分析、基因表达数据分析等），形成了数据挖掘领域。

模式分类是模式识别的主要任务和核心研究内容。分类器设计是在训练样本集合上进行优化（如使每一类样本的表达误差最小或使不同类别样本的分类误差最小）的过程，也就是一个机器学习过程。由于模式识别的对象是存在于感知信号中的物体和现象，它研究的内容还包括信号/图像/视频的处理、分割、形状和运动分析等，以及面向应用（如文字识别、语音识别、生物认证、医学图像分析、遥感图像分析等）的方法和系统研究。

本文简要回顾模式识别领域的发展历史和主要方法的演变，介绍模式识别理论方法研究的最新进展并分析未来的发展趋势。由于 Jain 等人的综述[3]已经全面介绍了 2000 年以前模式分类方面的进展，本文侧重于 2000 年以后的研究进展。

2. 历史回顾

现代模式识别是在 20 世纪 40 年代电子计算机发明以后逐渐发展起来的。在更早的时候，已有用光学和机械手段实现模式识别的例子，如在 1929 年 Gustav Tauschek 就在德国获得了光学字符识别的专利。作为统计模式识别基础的多元统计分析和鉴别分析[4]也在电子计算机出现之前提出来了。1957 年 IBM 的 C.K. Chow 将统计决策方法用于字符识别[5]。然而，“模式识别”这个词被广泛使用并形成一个领域则是在 20 世纪 60 年代以后。1966 年由 IBM 组织在波多黎各召开了第一次以“模式识别”为题的学术会议[6]。Nagy 的综述[7]和 Kanal 的综述[8]分别介绍了 1968 年以前和 1968-1974 的研究进展。70 年代几本很有影响的模式识别教材(如 Fukunaga [9], Duda & Hart [10])的相继出版和 1972 年第一届国际模式识别大会（ICPR）的召开标志着模式识别领域的形成。同时，国际模式识别协会（IAPR）在 1974 年的第二届国际模式识别大会上开始筹建，在 1978 年的第四届大会上正式成立。

统计模式识别的主要方法，包括 Bayes 决策、概率密度估计（参数方法和非参数方法）、特征提取（变换）和选择、聚类分析等，在 20 世纪 60 年代以前就已经成型。由于统计方法不能表示和分析模式的结构，70 年代以后结构和句法模式识别方法受到重视。尤其是付京荪（K.S. Fu）提出的句法结构模式识别理论在 70-80 年代受到广泛的关注。但是，句法模式识别中的基元提取和文法推断（学习）问题直到现在还没有很好地解决，因而没有太多的实际应用。

20 世纪 80 年代 Back-propagation (BP) 算法的重新发现和成功应用推动了人工神经网络研究和应用的热潮。神经网络方法与统计方法相比具有不依赖概率模型、参数自学习、泛化性能[1]良好等优点，至今仍在模式识别中广泛应用。然而，神经网络的设计和实现依赖于经验，泛化性能不能确保最优。90 年代支持向量机(SVM)的提出吸引了模式识别界对统计学习理论和核方法(Kernel methods)的极大兴趣。与神经网络相比，支持向量机的优点是通过优化一个泛化误差界限自动确定一个最优的分类器结构，从而具有更好的泛化性能。而核函数的引入使很多传统的统计方法从线性空间推广到高维非线性空间，提高了表示和判别能力。

结合多个分类器的方法从 90 年代前期开始在模式识别界盛行，后来受到模式识别界和机器学习界的共同重视。多分类器结合可以克服单个分类器的性能不足，有效提高分类的泛化性能。这个方向的主要研究问题有两个：给定一组分类器的最佳融合和具有互补性的分类器组的设计。其中一种方法，Boosting，现已得到广泛应用，被认为是性能最好的分类方法。

进入 21 世纪，模式识别研究的趋势可以概括为以下四个特点。一是 Bayes 学习理论越来越多地用来解决具体的模式识别和模型选择问题，产生了优异的分类性能[11]。二是传统的问题，如概率密度估计、特征选择、聚类等不断受到新的关注，新的方法或改进/混合的方法不断提出。三是模式识别领域和机器学习领域的相互渗透越来越明显，如特征提取和选择、分类、聚类、半监督学习等问题成为二者共同关注的热点。四是由于理论、方法和性能的进步，模式识别系统开始大规

[1] 泛化性能指分类器在测试样本（没有用于训练的新样本）上的分类正确率。提到“分类性能”时一般也是指泛化性能。

模地用于现实生活，如车牌识别、手写字符识别、生物特征识别等。

模式识别方法的细节可以参考一些优秀的教材，比如 Bishop (2006) [11], Fukunaga (1990)[12], Duda, Hart & Stork (2001)[13]等。

3. 模式识别研究现状

3.1 模式识别系统和方法概述

模式识别过程包括以下几个步骤：信号预处理、模式分割、特征提取、模式分类、上下文后处理。预处理通过消除信号/图像/视频中的噪声来改善模式和背景间的可分离性；模式分割是将对象模式从背景分离或将多个模式分开的过程；特征提取是从模式中提取表示该模式结构或性质的特征并用一个数据结构（通常为一个多维特征矢量）来表示；在特征表示基础上，分类器将模式判别为属于某个类别或赋予其属于某些类别的概率；后处理则是利用对象模式与周围模式的相关性验证模式类别的过程。

模式识别系统中预处理、特征提取（这里指特征度量的计算，即特征生成）和后处理的方法依赖于应用领域的知识。广义的特征提取包括特征生成、特征选择和特征变换（维数削减）[2]。后两个过程和分类器设计一样，需要在一个样本集上进行学习（训练）：在训练样本上确定选用哪些特征、特征变换的权值、分类器的结构和参数。

由于句法和结构模式识别方法是建立在完全不同于特征矢量的模式表示基础上且还没有得到广泛应用，本文与 Jain 等人[3]一样，主要关注统计模式识别（广义地，包括神经网络、支持向量机、多分类器系统等）的进展。

Bayes 决策是统计模式识别的基础。将模式表示为一个特征矢量 $\mathbf{x}$（多维线性空间中的一个点），给定 M 个类别的条件概率密度 $p(\mathbf{x}|\omega_i)$，i=1,...,M, 则模式属于各个类别的后验概率可根据 Bayes 公式计算：

$$p(\omega_i|\mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} = \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{\sum_{j=1}^{M} P(\omega_j)p(\mathbf{x}|\omega_j)},$$

其中 $P(\omega_i)$ 是第 i 类的先验概率。根据 Bayes 决策规则，模式 $\mathbf{x}$ 被判别为后验概率最大的类别（最小错误率决策）或期望风险最小的类别（最小代价决策）。后验概率或鉴别函数把特征空间划分为对应各个类别的决策区域。

模式分类可以在概率密度估计的基础上计算后验概率密度，也可以不需要概率密度而直接近似估计后验概率或鉴别函数（直接划分特征空间）。基于概率密度估计的分类器被称为生成模型(Generative model)，如高斯密度分类器、Bayes 网络等；基于特征空间划分的分类器又被称为判别模型（Discriminative model），如神经网络、支持向量机等。生成模型每一类的参数在一类的

[2] “特征提取”在很多时候就是指特征变换或维数削减，有时候也指从模式信号计算特征度量的过程（特征生成）。这就需要根据语言的上下文来判断它的意思。

训练样本上分别估计，当参数模型符合样本的实际分布或训练样本数比较少时，生成模型的分类性能优良。判别模型在训练中直接调整分类边界，以使不同类别的样本尽可能分开，在训练样本数较多时能产生很好的泛化性能。但是，判别模型在训练时每一类参数的估计要同时考虑所有类别的样本，因而训练的计算量较大。

3.2 概率密度估计

概率密度估计和聚类一样，是一个非监督学习过程。研究概率密度估计主要有三个意义：分类、聚类（分割）、异常点监测(Novelty detection)。在估计每个类别概率密度函数的基础上，可以用 Bayes 决策规则来分类。概率密度模型经常采用高斯混合密度模型(Gaussian mixture model, GMM)，其中每个密度成分可以看作是一个聚类。异常点监测又称为一类分类(One-class classification)，由于只有一类模式的训练样本，在建立这类模式的概率密度模型的基础上，根据相对于该模型的似然度来判断异常模式。

高斯混合密度估计常用的 Expectation-Maximization (EM)算法[14]被普遍认为存在三个问题：估计过程易陷于局部极值点，估计结果依赖于初始化值，不能自动确定密度成分的个数。对于成分个数的确定，提出了一系列的模型选择准则，如 Bayes 准则[15]、最小描述长度(MDL)、Bayesian Information Criterion (BIC)、Akaike Information Criterion (AIC)、最小消息长度(MML)等[16]。Figueiredo 和 Jain 在一个扩展的 EM 算法中引入密度成分破坏(Annihilation)机制[16]，可以达到自动确定成分个数的目的。Ueda 和 Ghahramani 提出一种基于变分 Bayes 的准则，并用分裂-合并算法进行估计自动确定成分个数[17]。分裂-合并算法还可以同时克服局部极值影响。

高斯混合密度用于高维数据时会造成密度函数的参数太多，用于分类时还会降低泛化性能。这个问题可以通过限制协方差矩阵（为对角矩阵或单位矩阵的倍数）、参数共享或特征降维来克服。在多类分类时，不同类别的概率密度要建立在相同的特征空间。如果对不同类别或不同密度成分提取不同的子空间，则要将子空间的密度函数反投影到原来的特征空间[18]。Moghaddam 和 Pentland 的概率密度模型是主成分分析(PCA)子空间内的混合高斯密度和补子空间中的高斯密度的结合[19]。

最近，Bouguila 等人提出一种新的混合密度形式：Dirichlet 混合密度[20][21]。Dirichlet 分布表示离散概率（介于 0 到 1 之间且和等于 1）的联合分布，可以用于直方图、和归一化特征矢量等的概率密度估计。Dirichlet 密度可以是非对称的，比高斯密度函数更为灵活，但计算也更复杂。Dirichlet 混合密度可以用类似于 EM 的随机优化算法进行估计，在模式分类和图像聚类等应用中取得了优异的性能[21]。

概率密度估计的另一种新方法是稀疏核函数描述（支持向量描述）[22][23]。Schölkopf 等人采用类似支持向量机的方法，用一个核特征空间的超平面将样本分为两类，使超平面外的样本数不超过一个事先给定的比例[22]。该超平面的函数是一个样本子集（支持向量）的核函数的加权平均，可以像支持向量机那样用二次规划算法求得。Tax 和 Duin 的方法是用核空间的一个球面来区分区

域内和区域外样本[23]，同样地可以用二次规划进行优化。

3.3 特征选择

特征选择和特征变换都是为了达到维数削减的目的，在降低分类器复杂度的同时可以提高分类的泛化性能。二者也经常结合起来使用，如先选择一个特征子集，然后对该子集进行变换。近年来由于适应越来越复杂（特征维数成千上万，概率密度偏离高斯分布）的分类问题的要求，不断提出新的特征选择方法，形成了新的研究热点[24]。

特征选择的方法按照特征选择过程与分类器之间的交互程度可以分为过滤式(Filter)、Wrapper[25]、嵌入式、混合式几种类型。过滤式特征选择是完全独立于分类器的，这也是最常见的一种特征选择方式，选择过程计算量小，但是选择的特征不一定很适合分类。在 Wrapper 方法中，特征子集的性能使用一个分类器在验证样本上的正确率来衡量，这样选择的特征比较适合该分类器，但不一定适合其他的分类器。由于在特征选择过程中要评价很多特征子集（子集的数量呈指数级增长），即使采用顺序前向搜索，Wrapper 的计算量都是很大的，只适合特征维数不太高的情况。Wrapper 的另一个问题是当训练样本较少时会造成过拟合，泛化性能变差。

嵌入式方法是在分类器的训练过程中包含了特征选择功能，因此跟 Wrapper 一样也是依赖于分类器的。一个经典的方法是 LASSO[26]。近来有代表性的两种嵌入式方法是稀疏支持向量机[27]和 Boosting 特征选择[28]。混合式特征选择结合不同的方法以实现更好的计算复杂性-分类性能的折衷，在初始特征数量非常大时经常使用，如[29]的方法在三个阶段先后用三种方法削减特征个数：过滤、聚类、组合式选择。过滤方法和 Wrapper 也经常结合使用。

特征选择领域大部分的研究工作都集中在过滤式方法。模式识别领域早期的工作多把关注点放在搜索策略上[30][31]，特征子集评价准则多采用基于高斯密度假设的距离准则，如 Fisher 准则、Mahalanobis 距离等。其实，特征子集的评价准则更为重要，当准则较好地衡量特征子集的可分性且比较稳定时，简单的搜索策略就能产生良好的分类性能。下面分析两类比较有代表性的特征评价方法：基于间隔(Margin)的方法和基于互信息的方法。

RELIEF[32]是一种被广泛引用的过滤式特征选择方法，基本思想是根据特征空间中每个样本在正确类别和不同类别中的最近邻距离之差迭代调整特征的权值。这两个距离之差即我们今天所说的间隔。不过 RELIEF 并没有对一个全局的目标函数进行优化。最近提出来的一种迭代 RELIEF（I-RELIEF）方法设计一种基于间隔的全局目标函数，用类似 EM 的算法对特征的权值进行优化[33]。另一种方法则对特征子集的空间中最近邻分类的间隔进行优化[34]。

特征选择的基本原则是选择类别相关（Relevant）的特征而排除冗余的特征。这种类别相关性和冗余性通常用互信息(Mutual information, MI)来度量。特征与类别之间的互信息很好地度量了特征的相关性，而特征与特征之间的互信细则度量他们之间的相似性（冗余性）。因此，基于互信息的特征选择方法一般遵循这样一种模式：在顺序前向搜索中寻找与类别互信息最大而与前面已选特征互信息最小的特征[35]。[36]中提出的条件互信息用来度量在一个已选特征的条件下另一个新

的候选特征对分类的相关性。[37]通过分析一种相关度，Symmetrical Uncertainty (SU)与特征的Markov blanket 之间的关系，设计一种快速的两步特征选择方法：先根据单个特征与类别之间的相关度选出相关特征，第二步对相关特征根据特征-类别相关度和特征-特征相关度进行筛选。

3.4 特征变换

特征变换也常被称为特征提取，指从原始信号经过变换得到特征量的过程。传统的线性变换方法主要有主成分分析(PCA)和线性鉴别分析(LDA)，后者又叫 Fisher 鉴别分析(FDA)。LDA 的子空间学习是有监督的，目的是使子空间中类间离散度(Sb)和类内离散度(Sw)的行列式之比达到最大。LDA 假设各类样本服从高斯分布且不同类的协方差矩阵相同，而且所有样本在总体上服从高斯分布。另外，LDA 提取的特征个数受到类别数的限制，而当训练样本数相对特征维数较小时，Sw 为奇异，会带来很多计算上的问题。

由于非高斯分布、小样本等问题的存在，特征变换也是近年来研究的一个热点，这方面的工作可以分为以下几个方向：(1)针对小样本的线性特征提取方法；(2)类内协方差矩阵不同时的异方差(Heteroscedastic)鉴别分析；(3)非高斯分布下的特征提取方法；(4)局部空间特性保持的特征提取方法；(5)非线性特征提取方法；(6)二维模式特征提取方法。

小样本学习的一个典型例子是图像分类，如果直接用图像中所有象素点的值作为特征量，矢量的维数非常高，而每一类的样本数又很少。克服 Sw 奇异性的一个直接方法是正则化(Regularized)鉴别分析[38]，通过矩阵平滑使 Sw 变得非奇异。Fisherface 方法则用 PCA 把特征维数从 D 降到 N-M（N 是样本数，M 是类别数）使 Sw 变得非奇异[39]。但是，Sw 的维数由 D 降到 N-M 会损失一些鉴别信息，而降到 N-1 维则不会有损失[40]。而这时 Sw 仍然是奇异的，就需要从 Sw 的零空间（对应本征值为 0)提取一些特征[41]。与一般的 LDA 方法先对 Sw 对角化然后对 Sb 对角化相反，一种 Direct LDA 方法先对 Sb 对角化后从变换后的 Sw 提取对应较小本征值的鉴别矢量[42]。

对于类别协方差矩阵不同的情况异方差鉴别分析方法（如[43]）可以得到比 LDA 更好的分类性能。对于非高斯分布或任意分布的情况，非参数鉴别分析[44]是提取鉴别特征的一个基本思路。由此发展起来的方法还包括基于决策边界的鉴别分析[45][46]。在不假设参数概率密度的情况下，也可以用分类性能准则直接对鉴别投影矢量进行优化，这样的准则如最小分类错误(MCE)[47]和特征与类别之间的互信息[48]。对于每类样本为多模态分布的情况可以采用基于混合高斯密度的鉴别分析[49][50]。

局部性保持特征提取方法借鉴了流形学习（如 LLE 和 Isomap）的思想，目的是在子空间中保持样本点之间的相邻关系。流形学习的问题是只对训练样本进行投影，要推广到测试样本就需要用一个参数模型或回归网络来表示投影的过程。He 等人提出的局部性保持投影(LPP)方法通过优化一个局部性保持准则来估计投影矢量，可转换为矩阵本征值分解问题[51]。Yan 等人提出一种基于样本邻近关系分析的特征提取的统一框架，称为嵌入图(Embedded graph)，并在此基础上提出一种新的鉴别分析方法[52]。LPP 是一种非监督学习方法，被推广到监督学习和核空间 [53]。另外，Isomap

流形学习方法也被推广到监督学习用于非线性特征提取[54]。

几乎所有的线性特征投影方法都可以推广到核空间。Schölkopf 等人最先将核函数引入 PCA，提出 Kernel PCA (KPCA)方法[55]。类似地，将核函数引入 Fisher 鉴别分析，提出了 Kernel FDA (KFDA)[56]。[57]对核空间中结合 PCA 降维和 FDA 特征提取进行了深入的分析并提出了有效的算法。核空间的特征提取方法还有 Kernel Direct LDA [58], Kernel LPP [53]等。

二维模式主成分分析(2D-PCA)[59]或鉴别分析(2D-LDA)[60][61]是近年提出的一种针对图像模式的特征提取方法。这类方法直接在图像矩阵上计算协方差（离散度）矩阵。该矩阵的维数等于图像的行数或列数，计算起来简便多了。另外，矩阵投影到每个本征矢量得到一个矢量，而不是一个值，这样得到的特征值个数也远远多于 LDA。在高维图像人脸识别实验中，2D-PCA 和 2D-LDA 的分类性能分别优于 PCA 和 LDA。二维变换方法实际上是基于图像行或列的变换方法[62]，即对每一行或每一列分别投影得到特征，可以推广到基于图像块的投影。

3.5 分类器设计

模式分类是模式识别研究的核心内容，迄今为止提出了大量的分类方法。Jain 等人把分类器分为三种类型：基于相似度（或距离度量）的分类器、基于概率密度的分类器、基于决策边界的分类器。第一种分类器的性能取决于相似度或距离度量的设计，同时也取决于标板(Prototype)的学习。标板学习有多种方法[63]，如聚类、LVQ (Learning Vector Quantization)、经验风险最小化等。LVQ 和经验风险最小化可以看作是决策边界调整的学习方法，而聚类的作用类似概率密度估计。因此，我们把分类器分为以下三类：生成模型（包括概率密度模型）、判别模型（决策边界学习模型）、混合生成-判别模型。由于前面已专门介绍了概率密度估计，下面我们着重介绍后两类分类器。

判别学习分类器包括我们熟知的人工神经网络、LVQ[63]、支持向量机(SVM)[64]、Boosting[65]等。其共同特点是在学习过程中最优化一个目标函数（准则），该函数表示训练样本集上的分类错误率、错误率的上界、或与分类错误率相关的损失，称为经验风险(Empirical loss)。常见的风险函数如神经网络中常用的平方误差函数、交叉熵(Cross entropy, 又称为对数损失或对数似然度)、最小分类错误准则(MCE)[66]、SVM 中的间隔（Margin）、Boosting 中的指数损失、以及最近常用的条件似然度(Conditional likelihood)等。指数损失即间隔的指数，是分类错误率的一个上界。对数损失也被用于 Boosting 学习，称为 LogitBoost[67]。条件似然度是训练样本正确类别后验概率的对数。在二类情况下，条件似然度等同于对数似然度。

混合生成-判别学习的模式识别方法近年来受到广泛的关注[68-71]。这种方法结合了生成模型和判别模型的优点：生成模型表示了类别的概率分布或内部结构、在少量样本上学习可得到较高的泛化性能、对噪声/异常模式具有抗拒性；判别模型在概率分布不容易用参数模型表示和训练样本较多时泛化性能优异，而且判别学习得到的模型一般较小（参数较少）。混合生成-判别学习的方法一般是先对每一类模式建立一个生成模型（概率密度模型或结构模型），然后用判别学习准则对

生成模型的参数进行优化[71]。学习的准则可以是生成模型学习准则（如最大似然准则）和判别学习准则（如条件似然度）的加权组合[72][73]。结合判别学习的 Bayes 网络（如[74][75]）也可以看作是混合生成-判别学习模型。

分类器模型和学习方法多种多样，性能各有特点。一般来说，SVM 和 Boosting 在大部分情况下分类性能优异，但也有他们自身的不足：SVM 的核函数选择和 Boosting 的弱分类器选择对性能影响很大，分类的计算复杂度较高（如 SVM 的支持向量个数往往很大）。

多分类器方法被认为是结合不同分类器的优点、克服单个分类器性能不足的一个有效途径。早期的多分类器研究主要集中在对给定多个分类器的有效融合[3]。由于单个分类器之间的发散性(Divergence)或互补性对融合后的分类性能起重要作用，近年来的研究主要集中在如何设计互补性强的分类器组(Ensemble 或 Committee)。这方面的方法主要有集成神经网络(Ensemble neural network)、 Bagging[76]、 Boosting[77]、随机子空间(Random subspace)[78]、特征选择法[79]、纠错输出编码(Error-correcting output codes, ECOC)[80]等。Bagging 和 Boosting 都是通过对训练样本集进行重采样或加权来训练多个分类器，不过 Bagging 是并行的，而 Boosting 是串行的。随机子空间法和特征选择法都是从原有的特征集选择多个特征子集，特征子集的独立性决定了分类器的互补性。

与其他学习方法对样本集或特征集进行分解不同的是，ECOC 是对类别集进行分解，通过组合多个二类分类器（这里的一类可以是一个类别子集）来实现多类分类。另外一种通过二类分类器实现多类分类的方法是把一对样本之间的关系分为“同类”(Intra-class)和“不同类”(Extra-class)两类，输入特征从两个样本提取（如两个样本对应特征的差），二类分类器的输出给出两个样本“同类”的概率或相似度，多类问题采用近邻规则进行分类。这种方法可以克服训练样本不足的问题，而且在训练后可任意增加或减少类别而不必重新训练，近年来已广泛用于人脸识别等生物特征识别问题[81]。

4. 发展趋势

除了上面介绍的最新研究进展，模式识别领域的前沿研究方向还有：Bayes 学习、半监督学习、弱监督学习等。Bayes 学习得到的分类器参数并不是一些固定值，而是参数的概率分布。参数的先验概率分布函数形式的选择、超参数（先验概率分布的参数）的确定在计算上是比较复杂的。在识别时，需要对分类器的参数进行随机采样，然后把很多个参数值得到的分类结果组合起来，因而识别的计算量也是很大的。近年来，基于 Bayes 学习的分类器设计取得了明显进展[11][82]等，得到了优异的分类性能。但是，这些方法的计算还是很复杂的，对于大类别数、大样本集的学习问题还难以实现。

在大部分应用情况下，模式分类器经过训练后就固定不变，或者使用相当长一段时间才重新训练一次。在训练分类器时，样本的数量和代表性总是不够的，这就希望分类器能不断地适应新的样本而不损失对原来训练过的样本的分类性能。这样的增量学习问题很早就受到关注，提出了很多具

体的方法（如[83][84]），但还没有一个统一的理论框架。新增加的样本可能是没有类别标记的，因为无标记样本很容易得到，而标记过程费时费力。同时对标记样本和无标记样本进行学习的过程称为半监督学习，这是近年来机器学习领域的一个研究热点[85]。在标记样本比较少的情况下采用无标记样本能有效提高完全监督学习的分类性能。

大多数模式识别问题假设模式是与背景信号和其他模式分离的且表示成一个特征矢量。实际上，模式的分割不是一件简单的事情，一个固定长度的特征矢量也不一定能最好地表示模式的特性。在实际应用问题中经常要将模式分类与分割问题统一考虑，有些模式被表示成结构性数据结构（如属性图、概率图）。这些方面出现了大量的研究工作，这里不打算细述。目前有一类广受关注的模式识别问题，识别对象是没有分割的图像，训练图像的标记是其中有没有某一类目标，而不知道目标的具体位置、大小和方位。对这种标记不足的样本进行训练和识别的方法可以统称为弱监督学习[86][87]，可用于目标识别、图像检索、景物分类等。

研究计算机模式识别的目的是让机器具备人的感知和认知能力，代替人完成繁重的信息处理工作。当我们把计算机的模式识别能力与人的模式识别（视觉、听觉感知）能力相比，就会发现现有的模式识别方法与人的感知过程有很大区别，在性能上也相差很远，很多对人来说是轻而易举的事情对计算机来说却很难做到。这是由于目前对人的感知过程的机理和大脑结构还不是很了解，即使已经了解的部分也不容易在计算上或硬件上模拟。进一步研究人的感知机理并借鉴该机理设计新的模式识别计算模型和方法是将来的一个重要方向。

5. 总结

本文围绕模式分类这个模式识别的核心问题概述了近年来在概率密度估计、特征选择和变换、分类器设计等方面的重要研究进展，并分析了最近的发展趋势。由于篇幅限制，对模式识别的其他问题，包括分割、上下文处理、计算机视觉、以及重要的应用领域（语音识别、文字识别、生物特征识别等）没有展开讨论。本文力求引用本领域最有代表性的研究工作，但难免会遗漏一些重要的工作，请有关作者谅解。

参考文献

[1] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review*, 65: 386-408, 1958.

[2] N.J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.

[3] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, *IEEE Trans. PAMI*, 22(1): 4-37, 2000.

[4] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7: 179-188, 1936.

[5] C.K. Chow, An optimum character recognition system using decision functions, *IRE Trans. Electronic Computers*, 6: 247-254, 1957.

[6] G. Nagy, Pattern Recognition 1966 IEEE Workshop, *IEEE Spectrum*, 92-94, Feb. 1967.

[7] G. Nagy, State of the art in pattern recognition, *Proc. IEEE*, 56(5): 836-862, 1968.

[8] L.N. Kanal, Patterns in pattern recognition: 1968-1974, *IEEE Trans. Information Theory*, 20(6): 697-722, 1974.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

[10] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

[11] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second edition, Academic Press, 1990.

[13] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, second edition, John Wiley & Sons, New York, 2001.

[14] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, *J. Royal Statistics Society B*, 39: 1-38, 1977.

[15] S.J. Roberts, D. Husmeier, L. Rezek, W. Penny, Bayesian approaches to Gaussian mixture modeling, *IEEE Trans. PAMI*, 20(11): 1133-1142, 1998.

[16] M.A.T. Figueiredo, A.K. Jain, Unsupervised learning of finite mixture models, *IEEE Trans. PAMI*, 24(3): 381-396, 2002.

[17] N. Ueda, Z. Ghahramani, Bayesian model search for mixture models based on optimizing variational bounds, *Neural Networks*, 15(10): 1223-1241, 2003.

[18] P.M. Baggenstoss, The PDF projection theorem and the class-specific method, *IEEE Trans. Signal Processing*, 51(3): 672-685, 2003.

[19] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, *IEEE Trans. PAMI*, 19(7): 696-710, 1997.

[20] N. Bouguila, D. Ziou, J. Vaillancourt, Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application, *IEEE Trans. Image Processing*, 13(11): 1533-1543, 2004.

[21] N. Bouguila, D. Ziou, A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture, *IEEE Trans. Image Processing*, 15(9): 2657-2668, 2006.

[22] B. Schölkopf, J. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, *Neural Computation*, 13(7): 1443-1471, 2001.

[23] D.M.J. Tax, R.P.W. Duin, Support vector data description, *Machine Learning*, 54(1): 45-66, 2004.

[24] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Machine Learning Research*, 3: 1157-1182, 2003.

[25] R. Kohavi, G. Tohu, Wrappers for feature selection, *Artificial Intelligence*, 97(1-2): 273-324, 1997.

[26] R. Tibshirani, Regression selection and shrinkage via the lasso, *J. Royal Statistical Society Series B*, 58(1): 267-288, 1996.

[27] J. Bi, K. Bennett, M. Embrecht, C. Breneman, M. Song, Dimensionality reduction via sparse support vector machines, *J. Machine Learning Research*, 3: 1229-1243, 2003.

[28] P. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, *Proc. CVPR 2001*, Hawaii, 2001, Vol.1, pp.511-518.

[29] J. Bins, B.A. Draper, Feature selection from huge feature sets, *Proc. 8th ICCV*, 2001, Vol.2, pp.159-165, 2001.

[30] A. Jain, D. Zongker, Feature selection: evaluation, application and small sample performance, *IEEE Trans. PAMI*, 29(2): 153-158, 1997.

[31] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition*, 33(1): 25-41, 2000.

[32] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, *Proc. ECML*, Catana, Italy, Springer-Verlag, 1994, pp.171-182.

[33] Y. Sun, Iterative RELIEF for feature weighting: algorithms, theories, and applications, *IEEE Trans. PAMI*, 29(6): 1035-1051, 2007.

[34] R. Gilad-Bachrach, A. Navot, N. Tishby, Margin based feature selection—theory and algorithms, *Proc. 21th ICML*, Alberta, Canada, 2004.

[35] R. Battiti, Using mutual information for selecting features in supervised neural net learning, *IEEE Trans. Neural Networks*, 5(4): 537-550, 1994.

[36] F. Fleuret, Fast binary feature selection with conditional mutual information, *J. Machine Learning Research*, 5: 1531-1555, 2004.

[37] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, *J. Machine Learning Research*, 5: 1205-1224, 2004.

[38] D.-Q. Dai, P.C. Yuen, Regularized discriminant analysis and its application to face recognition, *Pattern Recognition*, 36(3): 845-847, 2003.

[39] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class-specific linear projection space, *IEEE Trans. PAMI*, 19(7): 711-720, 1997.

[40] J. Yang, J.-y. Yang, Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36(2): 563-566, 2003.

[41] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, G.-J. Yu, A new LDA-based face recognition system which can solve the small sample size, *Pattern Recognition*, 33(10): 1713-1726, 2000.

[42] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data---with application to face recognition, *Pattern Recognition*, 34(10): 2067-2070, 2001.

[43] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, *IEEE Trans. PAMI*, 26(6): 732-739, 2004.

[44] K. Fukunaga, J. Mantock, Nonparametric discriminant analysis, *IEEE Trans. PAMI*, 5: 671-678, 1983.

[45] C. Lee, D.A. Landgrebe, Feature extraction based on decision boundary, *IEEE Trans. PAMI*, 15(4): 388-400, 1993.

[46] J. Zhang, Y. Liu, SVM decision boundary based discriminative subspace induction, *Pattern Recognition*, 38(10): 1746-1758, 2005.

[47] X. Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, *Pattern Recognition*, 36(10): 2429-2439, 2003.

[48] K.E. Hild II, D. Erdogmus, K. Tokkola, J.C. Principe, Feature extraction using information-theoretic learning, *IEEE Trans. PAMI*, 28(9): 1385-1392, 2006.

[49] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, *J. Royal Statist. Soc. Ser. B*, 58(1): 155-176, 1996.

[50] M. Zhu, A.M. Matinez, Subclass discriminant analysis, *IEEE Trans. PAMI*, 28(8): 1274-1286, 2006.

[51] X. He, S. Yan, Y. Hu, H.-J. Zhang, Learning a locality preserving subspace for visual recognition, *Proc. 9th ICCV*, Nice, France, 2003, pp.385-392.

[52] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extension: a general framework for dimensionality reduction, *IEEE Trans. PAMI*, 29(1): 40-51, 2007.

[53] J. Cheng, Q. Liu, H. Lu, Y.-W. Chen, Supervised kernel locality preserving projections for face recognition, *Neurocomputing*, 67: 443-449, 2005.

[54] X. Geng, D.-C. Zhan, Z.-H. Zhou, Supervised nonlinear dimensionality reduction for visualization and classification, *IEEE Trans. SMC Part B*, 35(6): 1098-1107, 2005.

[55] B. Schölkopf, A. Smola, K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10(5): 1299-1319, 1998.

[56] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, *Neural Computation*, 12(10): 2385-2404, 2000.

[57] J. Yang, A.F. Frangi, J.-y. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Trans. PAMI*, 27(2): 238-244, 2005.

[58] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Networks*, 14(1): 117-126, 2003.

[59] J. Yang, D. Zhang, A.F. Frangi, J.-y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, *IEEE Trans. PAMI*, 26(1): 131-137, 2004.

[60] J. Yang, D. Zhang, X. Yong, J.-y. Yang, Two-dimensional discriminant transform for face recognition, *Pattern Recognition*, 38(7): 1125-1129, 2005.

[61] J. Ye, R. Janardan, Q. Li, Two-dimensional linear discriminant analysis, *Advances in Neural Information Processing Systems*, 2004.

[62] L. Wang, X. Wang, X. Zhang, J. Feng, The equivalence of two-dimensional PCA to line-based PCA, *Pattern Recognition Letters*, 26(1): 57-60, 2005.

[63] C.-L. Liu, M. Nakagawa, Evaluation of prototype learning algorithms for nearest neighbor classifier in application to handwritten character recognition, *Pattern Recognition*, 34(3): 601-615, 2001.

[64] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[65] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1): 119-139, 1997.

[66] B.-H. Juang, S. Katagiri, Discriminative learning for minimum error classification, *IEEE Trans. Signal Processing*, 40(12): 3043-3054, 1992.

[67] J.H. Friedman, T. Hastie, T. Tibshirani, Additive logistic regression: a statistical view of boosting, *The Annals of Statistics*, 38(2): 337-374, 2000.

[68] Y.D. Rubenstein , T. Hastie, Discriminative vs informative learning, *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining*, Newport Beach, CA, 1997, pp.49-53.

[69] A.Y. Ng, M.I. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes, *Advances in Neural Information Processing Systems 14*, T.G. Dietterich, S. Becker, A. Ghahramani (eds.), MIT Press, 2002.

[70] I. Ulusoy, C.M. Bishop, Generative versus discriminative methods for object recognition, *CVPR 2005*,

Vol.2, pp.258-265.

[71] A. Holub, P. Perona, A discriminative framework for modeling object classes, *CVPR 2005*, Vol.1, pp.664-671.

[72] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, *IEEE Trans. Neural Networks*, 15(2): 430-444, 2004.

[73] J.A. Lasserre, C.M, Bishop, T.P. Minka, Principled hybrids of generative and discriminative models, *CVPR 2006*, New York, Vol.1, pp.87-94.

[74] D. Grossman, P. Domingos, Learning Bayesian network classifiers by maximizing conditional likelihood, *Proc. 21th ICML*, Alberta, Canada, 2004.

[75] R. Greiner, X. Su, B. Shen, W. Zhou, Structural extension to logistic regression: discriminative parameter learning of belief net classifiers, *Machine Learning*, 59(3): 297-322, 2005.

[76] L. Breiman, Bagging predictors, *Machine Learning*, 24(2): 123-140, 1996.

[77] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1): 119-139, 1997.

[78] T.K. Ho, The random subspace method for constructing decision forests, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(8): 832-844, 1998.

[79] D. Opitz, Feature selection for ensembles, *Proc. AAAI*, 1999.

[80] T.G. Dietterich, G. Baliri, Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research*, 2: 263-286, 1995.

[81] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, *Pattern Recognition*, 33(11): 1771-1782, 2000.

[82] M.A.T. Figueiredo, Adaptive sparseness for supervised learning, *IEEE Trans. PAMI*, 25(9): 1150-1159, 2003.

[83] R. Polikar, L. Udpa, A.S. Udpa, V. Honavar, Learn++: an incremental learning algorithm for supervised neural networks, *IEEE Trans. SMC—Part C: Applications and Review*, 31(4): 497-508, 2001.

[84] G. Cauwenberghs, T. Poggio, Incremental and decremental support vector machine learning, *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, V. Tresp (eds.), MIT Press, 2001.

[85] O. Chapelle, B. Scholkopf, A. Zien, *Semi-Supervised Learning*, The MIT Press, 2006.

[86] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, *Proc. IEEE Int. Conf. CVPR*, 2003, Vol.2, 264-271.

[87] D. Crandall, D. Huttenlocher, Weakly supervised learning of part-based spatial models for visual object recognition, *ECCV 2006*, LNCS 3979, Springer, 2006.

作者介绍

刘成林，中国科学院自动化研究所研究员、模式识别国家重点实验室副主任。1995 年在中国科学院自动化研究所获得工学博士学位。中国计算机学会高级会员。研究兴趣包括模式识别、机器学习、文字识别与文档分析等。

谭铁牛，中科院自动化所研究员，模式识别国家重点实验室主任。1989 年获英国帝国理工学院博士学位。中国图形图像学会常务副理事长、中国计算机学会与中国自动化学会副理事长、IEEE 与 IAPR Fellow。研究兴趣包括模式识别、计算机视觉、生物特征识别等。