统 计 模 式 识 别 的 研 究*

接　标，　刘冠晓，　冯乔生
（云南师范大学计信学院，云南 昆明 650092）

摘　要：　模式识别的基本目标就是在有监督和无监督的情况下进行分类，在模式识别的各种模型中，基于统计的方法已经被广泛的研究并应用于实际。一些新的技术和方法的引入（神经网络和支持向量机方法）也促进了该领域发展，文章研究了基于统计的模式识别的进展情况，给出了模式识别的概念、原理、方法并对各种方法进行了相关的评述。

关　键　词：　模式；模式识别；神经网络；特征选择；支持向量机
中图分类号：　TP391.4　　文献标识码：　A　　文章编号：　1007－9793(2005)06－0019－03

模式识别是人类的一项基本的能力，一个儿童可以非常容易的识别一个物体，识别图像、文字、数字等，然而这对计算机来说确是一项及其困难的任务，怎样让计算机去完成对完成这些任务，简言之，怎样让计算机具有人的智能，是当前人工智能研究的主要内容，而人工智能的研究的一个重要的子领域－模式识别则主要研究怎样让计算机能够观察环境，能从各种不同的背景中区分不同的模式，并且能够根据各种不同模式做出相应的合理的决策。在模式识别的各种模型中，基于统计的方法已被广泛的研究并应用于实际，神经网络技术和支持向量机的引入又极大促进了这块领域的发展，在过去 50 多年里该领域取得了很大的进展，但是仍存在一些复杂的模型问题仍然没有得到解决，并且新出现的应用（例如：数据挖掘、Web 挖掘，多媒体数据的检索、人脸识别、手写体识别）要求更健壮更有效率的模式识别技术。

1　模式、模式识别、模式识别系统的基本概念

Watanabe[1] 定义了一个模式是混沌世界的对立面，它是一个实体，并且有相应的名字，一个模式可以是指纹图像、手写汉字、人脸、语音信号等等。广义的说：存在于时间和空间上可观察的事物，如果可以区分他们是否相同或相似都可以称为模式；狭义的说：模式是通过对具体事物的观测所得到的具有时间和空间分布的信息。把模式所属的类别或同类中模式的总体称为模式类[2]。

模式识别是一门科学，它主要利用统计学、概率论、计算几何、机器学习、信号处理以及算法的设计等工具从可感知的数据中进行推理的一门学科。他的中心任务就是找出某“类”事物的本质属性，即在一定的度量和观测的基础上把待识别的模式划分到各自模式类中。给定一个模式对他们的识别/分类将面临两类任务：监督分类和无监督分类，其中有监督分类把模式划分已有的类别中，而无监督分类把模式划分到目前为止仍然不知的类别中。

模式识别系统主要由三部分组成：（1）数据的预处理：数据预处理是将输入模式的原始信息转换为利于计算机处理的数据；（2）特征的选择/特征的提取：选择模式的合理表示；（3）模式分类：利用训练样本集和已有的信息对计算机进行训练，从而制定出分类的标准，用于对待识别的模式进行分类。一个合理的分类系统通常要求具有从样本集或训练集中学习的能力。图 1 是一个典型的模式识别系统的详细框图[3]：传感器把图像声音等物理输入转换为输入信号；分割器把物体与背景及其它物体分开；特征提取器测量用于分类的物理属性；分类器根据特征给物体赋予类别标记；后处理器作其它考虑，例如上下文信息、错误代

* 收稿日期：2005－05－31
作者简介：接　标（1977－），男，安徽省宿州市人，硕士研究生，主要研究方向：模式识别与机器视觉.
通讯作者：冯乔生，男，云南省昆明市人，研究生导师，副教授，主要从事机器视觉与虚拟现实方向的研究..

价、选择合适的动作。并且大部分系统则采取了反馈机制(图中虚线)来完成一定的学习功能。

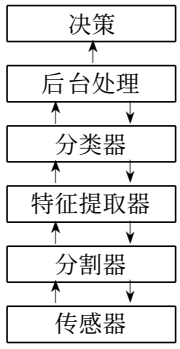

图 1 模式识别系统

Fig. 1 Pattern Recognition systems

2 模式分类的方法

有四种基本的模式分类方法[3]:(1)模板匹配的方法:它是一种相对简单和早期使用的方法,它的基本思想就是利用实体的特征(点线面及形状)进行模板匹配,通常模板本身具有从样本集中学习的功能,这种方法在实际中是可行的,但执行的效率不高,而且当模式由于图像处理,视觉点的改变或模式间内部类变化而损坏时,此方法将不再有效。(2)统计的方法:主要结合统计概率论的贝叶斯决策系统进行模式识别的技术,又称为决策理论识别方法,在该方法中,模式被表示为 d 维特征向量并且每个模式被看成 d 维特征向量空间的一个点。选择的特征应尽量使不同种类的模式位于 d 维特征向量空间中不相交的区域,而决策边界是由模式的概率分布决定的。(3)结构(句法)模式识别方法:利用模式与子模式分层结构的树状信息来完成的模式识别工作。其基本思想是把复杂的模式用简单的子模式或基元递归来描述,这种描述与语言中的句子通过单词来描述十分相似,因此形式语言中的许多方法都可以被应用,但句法方法在本质上是一种串操作,而且对本身具有噪音的模式进行分割时会遇到很多困难,这一性质使句法方法的应用带来很大的局限性。(4)神经网络方法:神经网络能被看作由大量的相互交互的简单神经元所构成平行的计算系统。人工神经网络主要模拟了人脑神经系统的工作的特点(例如:神经元的广泛连接、并行分布式信息的存储与处理、自适应学习等),虽没有人脑那么复杂,但它们之间存在许多相似之处(例如:两个网络的构成都是可计算单元的高度互连,并且处理单元之间的互连决定了网络的性能)。人工神经网络特别是向前反馈型网络由于具有自适应的学习能力而被广泛的应用到模式分类领域中,人工神经网络方法的引入也使得计算机视觉和模式识别处于一个新的迅速发展阶段。

3 统计模式识别

统计模式识别是目前最成熟也是应用最广泛的方法,它主要利用贝叶斯决策规则解决最优分类器问题。统计决策理论的基本思想就是在不同的模式类中建立一个决策边界,利用决策函数把一个给定的模式归入相应的模式类中。统计模式识别的基本模型如图 2[1],该模型主要包括两种操作模型:训练和分类,其中训练主要利用已有样本完成对决策边界的划分,并采取了一定的学习机制以保证基于样本的划分是最优的;而分类主要对输入的模式利用其特征和训练得来的决策函数而把模式划分到相应模式类中。

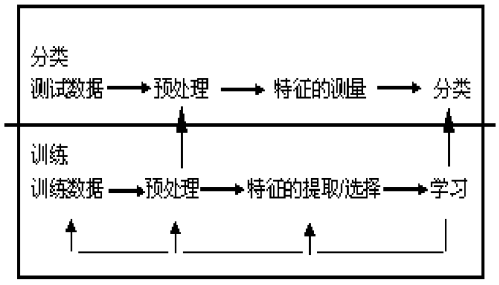

图 2 统计模式识别模型

Fig. 2 Model for statistical Pattern Recognition

令$\{w_1, w_2, \cdots, w_c\}$表示 C 个有限的类别集,向量 $\boldsymbol{X}$ 表示一个 d 维的特征向量,$p(\boldsymbol{X} \mid w_j)$ 表示 $\boldsymbol{X}$ 状态下条件概率密度函数,$P(w_j)$ 表示类别处于 w_j 时的先验概率,则贝叶斯定理可表示为:

$$p(w_j \mid X) = \frac{p(X \mid w_j)P(w_j)}{p(X)}$$

其中 $p(X) = \sum_{j=1}^{c} p(X \mid w_j)P(w_j)$

贝叶斯分类规则可描述为:对任何的 $i \neq j$ 如果 $p(w_i \mid X) > p(w_j \mid X)$ 则判为 w_i。

[4] 中已证明贝叶斯分类器在最小化分类错误率上是最优的。但有时错误率最小并不一定是最好的标准, 在考虑到不同的错误可能造成的后果的严重程度不同的情况下文献[5] 提出了基于最小风险的分类准则。

在统计模式识别中, 贝叶斯决策规则从理论上解决了最优分类器的设计问题, 但贝叶斯方法存在一个严重的缺点就是计算条件概率函数通常是非常困难的, 在实践中条件概率通常是未知的, 必须从可利用的样本集中估计得来, 然而在估计类条件概率时, 一方面在大部分情况下, 可利用的样本数总显得太少。另一方面当用于表示特征的向量的维数较大时会出现"维数灾难" 问题[3], 为了解决类条件概率的估计问题, 人们提出了各种解决方法:

3.1 最大似然估计和贝叶斯估计

这两种方法都是首先假定类条件概率密度的形式已知(例如高斯分布) 而参数未知(其中 μ、σ) 的情况下, 利用现有的样本对参数进行估计。参数估计问题是统计学中的经典问题, 最常用最有效的方法就是最大似然估计和贝叶斯估计, 最大似然估计把待估计的参数看作确定性的量, 只是其取值未知, 最大似然估计方法所要寻找的是能最好解释训练样本的那个参数值; 贝叶斯估计把待估计的参数看成是符合某种先验概率分布的随机变量, 而训练样本的作用就是把先验概率转化为后验概率。而递归贝叶斯方法通过逐次修正的办法来更新贝叶斯参数估计的结果。在实际中通常更多的使用最大似然估计方法, 因为该方法更容易实现, 并且在大样本的情况下, 得到的分类器的效果也较好。

3.2 非参数技术

上述的方法是在假定类条件概率密度的形式已知而参数未知的情况下利用样本完成对参数的估计。但如果类条件概率密度的形式也未知情况下, 我们只有或从样本中估计概率密度(例如: Parzen 窗方法), 或者直接利用样本来构建决策边界(例如 k 近邻方法)。其中 k 近邻方法的计算复杂度非常大, 通常人们使用部分距离、预建立结构、剪辑方法等手段来降低复杂度, 而且为了处理某些不变性的问题, Simard 等人提出了切空间距离的概念[6], 并且可以和 k 近邻方法相结合使用。

3.3 线性判断法

基于概率密度的方法要求首先利用样本对概率密度进行估计, 而后构建判别函数, 从而确定决策边界。但有些情况下我们可以直接利用样本通过最小化准则函数而直接构建决策边界, 在此情况下, 寻找线性判断函数问题可被形式化为寻找最小化准则函数问题。一系列的最小化准则函数的方法被提出[1]: 感知器最小化准则函数、松弛算法、Ho – Kashyap 算法等。[7] 很好的描述了线性判别函数在模式识别中应用, 它提出了最优化(最小化)线性判别函数问题并建议采用适当的梯度下降法从样本中求得解。20 世纪 90 年代中期, 统计学习理论和支撑向量机方法引起了广大研究人员的兴趣[8]。支撑向量机方法已在手写数字识别、文本分类等领域取得了良好的效果[9][10] 对该方法进行详细的论述。Fisher 判别法和主分量分析法也是传统线性方法, 主要用于特征抽取和模式分类。近年出现的基于核函数的 Fisher 判别法和基于核函数的主分量分析法是它们的线性推广。Minsky 和 Papert 的《感知器》一书指出线性分类器的弱点 – 但可以用神经网络的方法来解决。

4　总　结

模式识别从 20 世纪 20 年代发展至今, 人们的一种普遍看法是不存在对所有模式识别问题都适用的单一模型和解决所有识别问题的单一技术。早期, 统计模式识别研究的主要热点集中在贝叶斯决策理论、概率密度估计、"维数灾难" 问题和误差估计等。自从 90 年代初期统计模式识别经历了一个迅速发展的时期, 这主要由于新方法得引入(包括神经网络、机器学习、计算机科学等) 和新出现的应用(包括数据挖掘、文档分类等)。现在我们拥有了解决各类分类问题的方法, 在实际的分类中我们只要针对于不同的问题把各种方法结合起来, 取长补短, 推进模式识别的更大发展。

(下转第 38 页)

A theoretical study on the structure of Qinghaosu molecule

LIU You de[1], LI Xi ping[2], CHEN Xiu min[2]
(1. Yunnan Normal University, Kunming 650092, China;
2. Kunming University of science and Technology, Kunming 650093, China)

ABSTRACT: A theoretical study on geometric structure of the Qinghaosu molecule has been Carried out by ab initio method of the Quantum Chemistry. This research has been shown that whole molecular configuration, bond distances, bond angles and dihedral angle C – O – O – C etc quite the approximate to X – ray analytical volue, the result show ab initio method can be used in optimization of Qinghaosu and its derivations.

KEY WORDS: Qinghaosu; ab initio; sto – 3g basic set; 3 – 21g basic set; 6 – 31g basic set

---

(上接第 21 页)

参 考 文 献:

[1] Richard O. Duda Peter E. Hart David G. Stork. . 模式分类(第二版)[M]. 李宏东, 姚天翔译. 北京: 机械出版社, 2003.
[2] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000.
[3] Jain A K, Duin R P W, Jianchang Mao. Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4 ~ 37.
[4] Sergios Theodoridis, Konstantinos Koutroumbas. 模式识别(第二版)[M]. 李晶皎译. 北京: 电子工业出版社, 2004.
[5] Abraham wald. Contributions to the theory of statistical estimation and testing of hypotheses. Annals of mathematical Statistics, 10: 299 – 326, 1939.
[6] Patrice Simard, Yann Le Cun, and John Denker. Efficent pattern recognition using a new transformation distance. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, Advances inNeural Information Processing Systems, volume 5, pages 50 – 58, Morgan Kaufmann, SanMateo, CA 1993.
[7] Wilbur. H. Highleyman. Linear decision function, with application to pattern recognition. Procedding of the IRE, 50: 1501 – 1504, 1962.
[8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报. 2000, 126(11): 2000.
[9] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程. 2004, 30(10): 2004.
[10] Vladimir Cherkassky and Filip Mulier. Learning from data: Concepts, Theory, and Methods. Wiley, New York. 1998.

Research of Statistical Pattern Recognition

JIE Biao, LIU Guan xiao, FENG Qiao sheng
(Department of Computer Science and Information Technology,
Yunnan normal university, Kunming 650092, China), China)

ABSTRACT: the primary goal of pattern recognition is supervised or unsupervised classification. Among the various models, the statistical approach has been intensively studied and used in practice. More recently, new techniques and methods (neural network and support vector machine) boost developments of statistical pattern recognition. This paper researches recent developments of statistical pattern recognition, and give the concepts、principal 、methods of recognition and some comments about these methods.

KEY WORDS: pattern; pattern recognition; neural network; feature selection; support vector machine