新闻出版改革发展项目库入库项目

“十二五”国家重点图书

·藏文信息处理技术·

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

བོད་ཡིག་དབེ་དབྱིབས་ངོས་འཛིན་ལག་རྩལ་དང་བརྒྱུད་ཚོངས་ལག་བསྟར།

藏语模式识别技术及工程实践

欧　珠◎著

SWJTUP 西南交通大學出版社

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

新闻出版改革发展项目库入库项目

“十二五”国家重点图书

西藏大学博士学位授权立项建设项目

藏文信息处理技术

བོད་ཡིག་དེའི་དབྱིབས་ངོས་འཛིན་ལག་རྩལ་དང་བཀྲུད་ཉོངས་ལག་བསྟུར།

藏语模式识别技术及工程实践

欧　珠　著

西南交通大学出版社
·成　都·

内容简介

本书介绍了模式识别研究中的一些基本理论以及相关的模型，包括贝叶斯决策、线性判别函数、神经网络理论、隐马尔科夫模型、聚类技术等。重点结合藏语文模式识别实际问题，如印刷体藏文字符识别、木刻经文藏文字符识别、藏语语音识别等技术内容，从不同的研究角度介绍了这些问题的解决思路。

本书可作为自动化、计算机、电子和通信等专业研究生和高年级本科生的教材，也可作为学习模式识别和人工智能的参考资料，还可作为藏文信息处理技术研究生的教学参考书。

图书在版编目（CIP）数据

藏语模式识别技术及工程实践 / 欧珠著.
—成都：西南交通大学出版社，2015.3
（藏文信息处理技术）
ISBN 978-7-5643-3644-8

Ⅰ. ①藏… Ⅱ. ①欧… Ⅲ. ①藏语－文字识别－模式识别 Ⅳ. ①H214②TP391.4

中国版本图书馆 CIP 数据核字（2014）第 306940 号

藏文信息处理技术

藏语模式识别技术及工程实践

Zangyu Moshi Shibie Jishu ji Gongcheng Shijian

欧 珠 著

*

责任编辑 李芳芳
封面设计 墨创文化
西南交通大学出版社出版发行
四川省成都市金牛区交大路 146 号 邮政编码：610031
发行部电话：028-87600564
http://www.xnjdcbs.com
四川森林印务有限责任公司印刷

*

成品尺寸：210 mm × 285 mm 印张：13
字数：316 千字
2015 年 3 月第 1 版 2015 年 3 月第 1 次印刷
ISBN 978-7-5643-3644-8
定价：38.00 元

前　言

我们在日常生活中经常进行模式识别的活动，这是人类的一项基本智能。比如说，我们能够辨别桌子、椅子，很小的时候能够辨别出自己的父母，能够辨别出是谁的声音，能够进行正常的阅读。这些都是我们习以为常的能力。

模式识别（Pattern Recognition）是指对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。模式识别又常称作模式分类，随着20世纪40年代计算机的出现以及50年代人工智能的兴起，人们也希望能用计算机来代替或扩展人类的部分脑力劳动，所以模式识别在20世纪60年代初迅速发展并成为一门新学科。

模式识别和人工智能，在大学计算机、自动化、电子信息等专业中都是比较重要的课程。在学习和应用这些课程的时候，笔者深深体会到，有些算法，理解起来容易，但在实际应用时，却未必是很简单的事情，对于初学者而言，甚至一些简单问题也根本无法下手。

作者根据自己近年来承担的有关国家项目，参考了大量的模式识别与人工智能方面的书籍，在原有算法的基础上，结合藏语文特点，编写了印刷体藏文字符的识别、木刻经文藏文字符的识别、藏语拉萨语的语音识别等领域的一些实用算法，这些算法和程序供读者在学习中使用，请勿用于其他目的。对很多专著中已有详细阐述的经典算法，本书不再多述，只是简单地提及或给出参考文献，避免与其他专著在内容上过多的重复。

本书大部分程序是用VC++编写的，在VC环境中可以直接应用。

本书是在国家自然科学基金项目（编号：60863013，名称：木刻藏文经书识别系统中的特征提取算法的研究；编号：61250012，名称：面向拉萨语的自动发音错误检测方法研究）、科技部973计划前期研究专项项目（编号：2009CB326201，名称：藏语语音识别技术研究）、教育部高等学校科技创新工程重大项目培育资金

项目（编号：706059，名称：藏文文字识别技术研究及其实现）、藏文信息技术教育部创新团队（编号：IRT0975，名称：藏文信息技术“长江学者和创新团队发展计划”创新团队）、西藏自治区学术技术带头人及西藏大学珠峰学者等资助下所完成的项目成果之一，特此表示感谢。

本书由西藏大学藏文信息技术国家地方联合工程中心策划，由欧珠编写。此外，参加资料收集和整理的人员有赵栋才、扎西加、普次仁、裴春宝等。读者在阅读中，一旦发现书中存在难以理解之处，请查阅原始文献对照理解。由于时间仓促，加之经验不足，书中难免存在不妥之处，敬请读者批评指正，欢迎写信联系 ngodrup@126.com。

最后深深感谢曾经提供过源代码的朋友以及其他一些在作者学习过程中给予帮助的朋友。

欧 珠

2014 年 12 月

目　录

第 1 章

绪　论

模式识别诞生于 20 世纪 20 年代，随着 40 年代计算机的出现以及 50 年代人工智能的兴起，模式识别在 60 年代初迅速发展成为一门学科。它所研究的理论和方法在很多科学和技术领域中得到了广泛的重视，推动了人工智能系统的发展，扩大了计算机应用的可能性。为了使读者更好地掌握后面各章的内容，对这些内容的有效性和局限性有较全面的认识，正确地使用这些理论和方法，进而研究藏语模式识别技术的理论和方法，本章主要讨论和介绍模式识别的基本概念和问题，以对模式识别的现状与未来的发展方向有更全面的了解。

1.1　模式和模式识别的概念

模式识别是人类的一项基本智能，在日常生活中，人们经常通过“模式识别”来认识外界事物。随着 20 世纪 40 年代计算机的出现以及 50 年代人工智能的兴起，人们当然也希望能用计算机来代替或扩展人类的部分脑力劳动。模式识别在 20 世纪 60 年代初迅速发展并成为一门新学科。

什么是模式和模式识别？模式识别一词的英文是 Pattern Recognition。英文 Pattern 主要有两种含义，一是代表事物（个体或一组事物）的模板或原型，二是表征事物的特征或性状的组合。在模式识别学科中，模式可以看作是对象的组成成分或影响因素间存在的规律性关系，或者是因素间存在确定性或随机性规律的对象、过程或事件的集合。广义地说，存在于时间和空间中可观察的事物，如果可以区别它们是否相同或相似，都可以称之为模式；狭义地说，模式是通过对具体的个别事物进行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一类中模式的总体称为模式类（或简称类）。而“模式识别”则是在某些一定量度或观测基础上把待识模式划分到各自的模式类中去。

模式识别，直观而无所不在。“物以类聚，人以群分”。周围物体的认知：唐卡、地毯；人的识别：扎西、卓玛；声音的辨别：飞机、汽车、火车、狼叫；气味的分辨：炒青稞、炖牛肉。人和动物的模式识别能力是极其简单的，但对计算机来说却是非常困难的。

模式识别的研究主要集中在两方面，即研究生物体（包括人）是如何感知对象的，以及

在给定的任务下，如何用计算机实现模式识别的理论和方法。前者是生理学家、心理学家、生物学家、神经生理学家的研究内容，属于认知科学的范畴；后者属于数学家、信息学专家和计算机科学工作者的研究内容，通过近几十年来的努力，已经取得了系统的研究成果。

一个计算机模式识别系统基本上是由三个相互关联而又有明显区别的过程组成的，即数据生成、模式分析和模式分类。数据生成是将输入模式的原始信息转换为向量，成为计算机易于处理的形式。模式分析是对数据进行加工，包括特征选择、特征提取、数据维数压缩和决定可能存在的类别等。模式分类则是利用模式分析所获得的信息，对计算机进行训练，从而制定判别标准，对待识别模式进行分类。

模式识别有两种基本的方法，即统计模式识别方法和结（句法）模式识别方法。统计模式识别是对模式的统计分类，即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。利用模式与子模式分层结构的树状信息所完成的模式识别工作，就是结构模式识别或句法模式识别。

模式识别的应用包括文字识别、语音识别、图像识别、指纹识别等。

1.2 模式识别的发展历史

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已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，也在电子计算机出现之前被提出。1957年IBM的C. K. Chow将统计决策方法用于字符识别。然而，“模式识别”一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。Nagy的综述和Kanal的综述分别介绍了1968年以前和1968—1974年间的研究进展。70年代几本很有影响的模式识别教材，如Fukunaga，Duda & Hart的相继出版以及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，国际模式识别协会在1974年IAPR的第二届国际模式识别大会上开始筹建，于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）问题直到现在还未得到很好的解决，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的研究和应用热潮。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、聚类分析性能良好等优点，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。同时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。

21世纪以来，模式识别研究的趋势可以概括为以下四个特点。一是贝叶斯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。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。

1.3 模式识别的方法

模式识别与很多学科都有联系，如统计学、心理学、语言学、计算机科学、生物学、控制论等。它与人工智能、图像处理的研究也有交叉关系。例如自适应或自组织的模式识别系统包含了人工智能的学习机制，同时人工智能研究的景物理解、自然语言理解也包含模式识别问题。又如模式识别中的预处理和特征抽取环节应用图像处理的技术，图像处理中的图像分析也应用模式识别的技术。

模式识别的方法主要有决策理论方法和结构（句法）方法，模式识别方法的选择取决于问题的性质。如果被识别的对象极为复杂，而且包含丰富的结构信息，一般采用句法方法；被识别对象不是很复杂或不含明显的结构信息，一般采用决策理论方法。这两种方法不能截然分开，因为在句法方法中，基元本身就是用决策理论方法抽取的。在应用中，将这两种方法结合起来分别施加于不同的层次，常能收到较好的效果。

统计模式识别方法是对模式的统计分类，即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。利用模式与子模式分层结构的树状信息所完成的模式识别工作，就是结构模式识别或句法模式识别。统计模式识别的基本原理是：具有相似性的样本在模式空间中互相接近，并形成“集团”，即“物以类聚”。其分析方法是根据模式所测得的特征向量 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}(x\in\mathbf{R}^n)$，将一个给定的模式归入 C 个类 $\omega_1,\omega_2,\cdots,\omega_C$ 中，然后根据模式之间的距离函数来判别分类。

统计模式识别的主要方法有：判别函数法、K 近邻分类法、非线性映射法、特征分析法、主因子分析法等。

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，但其实施却必须首先解决更困难的概率密度估计问题。BP神经网络方法直接从观测数据（训练样本）学习，更为简便有效，因而获得了广泛的应用，但作为一种启发式技术，它缺乏指定工程实践的坚实理论基础。统计推断理论研究所取得的突破性成果促成现代统计学习理论——VC理论的建立，该理论不仅在严格的数学基础上圆满地解决了人工神经网络中出现的理论问题，而且导出了一种新的学习方法——支撑向量机（SVM）。

结构模式识别方法通过考虑识别对象的各部分之间的联系来达到识别分类的目的。识别采用结构匹配的形式，通过计算一个匹配程度值（matching score）来评估一个未知的对象或未知对象某些部分与某种典型模式间的关系。当成功地制定出了一组可以描述对象部分之间关系的规则后，可以应用一种特殊的结构模式识别方法——句法模式识别，来检查一个模式基元的序列是否遵守某种规则，即句法规则或语法。

神经网络是受人脑组织的生理学的启发而创立的，由一系列互相联系的、相同的单元（神经元）组成。相互间的联系可以在不同的神经元之间传递增强或抑制信号。这种增强或抑制是通过调整神经元相互间联系的权重系数（weight）来实现的。

神经网络可以实现监督和非监督学习条件下的分类。

1.4 模式识别系统的组成

模式识别有两种基本的方法，即统计模式识别方法和结构（句法）模式识别方法，与此相应的模式识别系统都是由两个过程（设计与实现）所组成。“设计”是指对一定数量的样本（训练集/学习集）进行分类设计器的设计。“实现”是指用所设计的分类器对待识别的样本进行分类决策。本书的例子主要是用统计模式识别方法，在用到结构模式识别的方法时，我们会对其再加以介绍。基于统计模式识别方法的系统主要由以下几个部分组成：信息获取、预处理、特征提取和选择、分类决策，如图 1-1 所示。

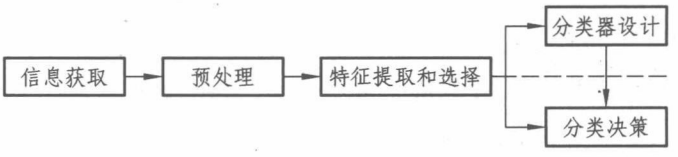

图 1-1 模式识别系统的基本组成

下面，我们对这几个部分作一个简要的说明。

1.4.1 信息获取

为了使计算机能够对各种现象进行分类识别，需要用计算机可以运算的符号来表示所研究的对象，通常输入对象信息有下列 3 种类型：

- 二维图像：如文字、照片、指纹、地图等这类对象；
- 一维波形：如脑电图、LI 电图、机械振动波形等；
- 物理参量和逻辑值：体温、化验数据、参量正常与否的描述。

1.4.2 预处理

预处理的目的是去除噪声、加强有用的信息，并对输入测量仪器或其他因素所造成的退

化现象进行复原。主要指图像处理，包括 A/D、二值化、图像的平滑、变换、增强、恢复、滤波等。

1.4.3 特征提取和选择

由图像或波形所获得的数据量是相当大的，例如，一个文字图像可以有几千个数据，一个心电图波形也可能有几千个数据，一个卫星遥感图像的数据量更大。为了有效地实现分类识别，要对原数据进行变换，得到最能反映分类本质的特征。这就是特征提取和选择的过程。

1.4.4 分类器设计

分类器设计的主要功能是通过训练确定判决规则，使按此类判决规则分类的错误率最低，并把这些判决规则建成标准库。

1.4.5 分类决策

在样本训练集基础上确定某个判决规则，使按这种判决规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小。

1.5 模式识别的应用

模式识别与很多学科都有联系，它与统计学、心理学、语言学、计算机科学、生物学、控制论等都有关系。它与人工智能、图像处理的研究也有交叉关系。例如自适应或自组织的模式识别系统包含了人工智能的学习机制，人工智能研究的景物理解、自然语言理解也包含模式识别问题。又如模式识别中的预处理和特征抽取环节应用图像处理的技术，图像处理中的图像分析也应用模式识别的技术。以下仅从几个常用的方面进行介绍。

1.5.1 文字识别

汉字已有数千年的历史，是世界上使用人数最多的文字，对于中华民族灿烂文化的形成和发展有着不可磨灭的功勋。所以在信息技术及计算机技术日益普及的今天，如何将文字方便、快速地输入计算机中已成为影响人机接口效率的一个重要因素，也关系到计算机能否真正在我国得到普及和应用。目前，汉字输入主要分为人工键盘输入和机器自动识别输入两种。其中人工键入速度慢且劳动强度大，自动输入又分为汉字识别输入及语音识别输入。从识别技术的难度来说，手写体识别的难度高于印刷体，而在手写体识别中，脱机手写体的难度又

远远超过了连机手写体识别。到目前为止，除了数字的脱机手写体识别已有实际应用外，汉字等文字的脱机手写体识别都还处在实验阶段。

1.5.2 语音识别

语音识别技术技术所涉及的领域包括信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等。近年来，在生物识别技术领域中，声纹识别技术以其独特的方便性、经济性和准确性等优势受到世人瞩目，并日益成为人们日常生活和工作中重要且普及的安全验证方式。而且利用基因算法训练连续隐马尔柯夫模型的语音识别方法现已成为语音识别的主流技术，该方法识别速度较快且有较高的识别率。

1.5.3 指纹识别

我们手掌及手指、脚、脚趾内侧表面的皮肤凹凸不平产生的纹路会形成各种各样的图案。而这些皮肤的纹路在图案、断点和交叉点上各不相同，是唯一的。依靠这种唯一性，就可以将一个人同他的指纹对应起来，通过将他的指纹和预先保存的指纹进行比较，便可以验证他的真实身份。一般的指纹分为以下几个大的类别：left loop，right loop，twin loop，whorl，arch 和 tented arch，这样就可以将每个人的指纹分别归类，进行检索。指纹识别基本上可分成预处理、特征选择和模式分类几个大的步骤。

1.6 本书的内容安排和程序

本书介绍了模式识别中的一些基本理论，在借鉴其他语言模式识别技术的理论、方法和技术的基础上，给出了藏语模式识别的一些应用实例，具体安排如下：

第 2 章介绍了模式识别中的一些基本决策方法，包括贝叶斯决策、线性判别函数等；第 3 章介绍了藏语及藏语信息处理；第 4 章介绍了藏语识别系统中的常用搜索算法；第 5 章介绍了印刷体藏文文字识别技术；第 6 章介绍了常见木刻经文文字识别技术；第 7 章介绍了藏语语音识别技术。本书对每一个实例，均给出了较为详细的分析过程，有的提供了多种方法，并给出了实现代码，具有一定的可学习性。

第 2 章

统计模式识别方法

第一章所介绍的模式识别一般有两种基本的识别方法，即统计模式识别方法（statistic pattern recognition）和结构（句法）模式识别方法（structure pattern recognition），与此相应的模式识别系统都由两个过程（设计与实现）所组成。本书重点考虑“统计模式识别”方法。“设计”是指对一定数量的样本（训练集/学习集）进行分类设计器的设计。“实现”是指用所设计的分类器对待识别的样本进行分类决策。

统计模式识别（statistic pattern recognition）的基本原理是：具有相似性的样本在模式空间中互相接近，并形成“集团”，即“物以类聚”。其分析方法是根据模式所测得的特征向量 $X_i=(x_{i_1},x_{i_2},\cdots,x_{i_d})^{\mathrm{T}}(i=1,2,\cdots,N)$，将一个给定的模式归入 C 个类 $\omega_1,\omega_2,\cdots,\omega_C$ 中，然后根据模式之间的距离函数来判别分类。其中，T 表示转置；N 为样本点数；d 为样本特征数。

统计模式识别的主要方法有：判别函数法、近邻分类法、非线性映射法、特征分析法、主因子分析法等。

2.1 分类与聚类

- Classification（分类），对于一个 classifier，通常需要你告诉它“这个东西被分为某某类”这样一些例子。理想情况下，一个 classifier 会从它得到的训练集中进行“学习”，从而具备对未知数据进行分类的能力，这种提供训练数据的过程通常叫作 supervised learning（监督学习）。

分类作为一种监督学习方法，要求必须事先明确地知道各个类别的信息，并且确定所有待分类项都有一个类别与之对应。但是很多时候上述条件得不到满足，尤其是在处理海量数据的时候，如果通过预处理使得数据满足分类算法的要求，则代价非常大，这时候可以考虑使用聚类算法。

常用的分类算法包括：决策树分类法、朴素的贝叶斯分类算法（native Bayesian classifier）、基于支持向量机（SVM）的分类器、神经网络法、K-最近邻法（K-Nearest Neighbor，KNN）、模糊分类法等。

- Clustering（聚类），对一批没有标出类别的模式样本集，按照样本之间的相似程度分

类，相似的归为一类，不相似的归为另一类，这种分类称为聚类分析，也称为无监督分类。简单地说，我们并不关心某一类是什么，我们需要实现的目标只是把相似的东西聚到一起，因此，一个聚类算法通常只需要知道如何计算相似度就可以开始工作了，所以 clustering 通常并不需要使用训练数据进行“学习”，这在 machine learning 中被称作 unsupervised learning（无监督学习）。

常见的聚类算法包括：*K*-均值聚类算法、*K*-中心点聚类算法、CLARANS、BIRCH、CLIQUE、DBSCAN 等。

2.1.1 贝叶斯定理

根据维基百科上的介绍，贝叶斯（Bayes）定理是关于随机事件 A 和 B 的条件概率和边界概率的一则定理，即

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

式中，$P(A|B)$是在 B 发生的情况下 A 发生的可能性。在贝叶斯定理中，每个名词都有约定俗成的名称。

- $P(A)$是 A 的先验概率或边界概率。之所以称为“先验”，是因为它不考虑任何 B 方面的因素。
- $P(A|B)$是已知 B 发生后 A 的条件概率（直白来讲，就是先有 B 而后才有 A），也因得自 B 的取值而被称作 A 的后验概率。
- $P(B|A)$是已知 A 发生后 B 的条件概率（直白来讲，就是先有 A 而后才有 B），也因得自 A 的取值而被称作 B 的后验概率。
- $P(B)$是 B 的先验概率或边界概率，也作标准化常量（normalized constant）。

因为 $P(B)$可以看作普通常量，所以我们只关心在给定事件 B 的情况下可能发生事件 A 的概率，$P(B)$的值是确定不变的，故有：

$$\arg\max_A \frac{P(B|A)P(A)}{P(B)}=\arg\max_A P(B|A)P(A)$$

由此，贝叶斯定理可表述为：后验概率 =（相似度 × 先验概率）/ 标准化常量，也就是说，后验概率与先验概率和相似度的乘积成正比。另外，比例 $\frac{P(B|A)}{P(B)}$ 有时也被称作标准相似度（standardised likelihood），Bayes 定理可表述为：后验概率 = 标准相似度 × 先验概率。

2.1.2 朴素贝叶斯算法

朴素贝叶斯算法是基于统计理论的方法，它能够预测类别所属的概率。朴素贝叶斯分类器假设一个指定类别中各属性的取值是相互独立的。这一假设也称为类别条件独立（class conditional independence），它可以有效地减少在构造朴素贝叶斯分类器时所需要的计算量。