模式识别教材总结

本教材共分为十章，主要围绕特征提取与选择，分类器设计和聚类分析，以及分类器和聚类结果的性能评价方法的主要内容讲述模式识别的基本概念、基础理论和典型方法，为从事模式识别研究的学者提供一本基础的教材，同时也为在其他学科应用模式识别技术的学者提供了基础。

第一章主要讲述了模式识别的基本概念和模式识别系统的基本构成，同时也展示了模式识别广阔的应用空间。

第二章主要讲述了以贝叶斯决策为核心的统计决策的基本思想和原理，介绍了最小错误率贝夜斯决策和在控制一类错误率的情况下使另一类错误率尽可能小的 Neyman-Pearson 决策的方法，还给出了正态分布下部分决策的具体形式和错误率的计算，并举例说明了离散情况下使用马尔可夫模型的统计决策方法。统计决策的基本原理就是根据各类特征的概率模型来估算后验概率，通过比较后验概率进行决策。而通过贝叶斯公式，后验概率的比较可以转化为类条件概率密度的比较，离散情况下也是类条件概率的比较，而这种条件概率或条件密度则反映了在各类的模型下观察到当前样本的可能性或似然度，因此可以定义两类之间的似然比或对数似然比进行决策。根据面对的具体问题不同，各类特征的概率模型可能会变得非常复杂，但是基本的求解步骤和决策原理是一致的。在概率模型准确的前提下，统计决策可以得到最小的错误率或者最小的风险，或者是实际问题中期望得到的两类错误率间最好的折中。因此，要使用统计决策进行模式识别，概率模型的估计问题就变得至关重要。

第三章主要讲述了条件概率密度的估计问题。然后又分章节介绍了最大似然估计的基本原理，求解方法和正态分布下的最大似然估计，然后又用贝叶斯估计的方法解决条件概率密度。最后又讲了

概率密度估计的非参数方法像 k_n 邻近估计方法和 Parzen 窗法的基本原理和算法。

第四章主要讲述了线性分类器。首先介绍了线性判别函数的基本概念和 Fisher 线性判别分析的原理方法。线性判别函数是形式最简单的判别函数。它虽然算法简单，但是在一定条件下能够实现或逼近最优分类器的性能，因此在很多实际问题中得到了广泛的应用。然后又介绍了感知器的概念和原理还有最小平方误差判别的方法，最后讲述了最优分类超平面与线性支持向量机的原理方法和多类线性分类器。

第五章讲述了几种非线性分类器的概念和设计方法，包括经典的分段线性分类器、二次判别函数和新近发展的多层感知器神经网络方法及支持向量机方法。然后又重点介绍了多层感知器方法和支持向量机方法是两种适用性很广的非线性方法，他们没有直接对非线性模型进行假设，而是通过多个隐节点作用的组合或样本与多个支持向量的核函数内积加权和来实现各种非线性分类面。在多层感知器中，分类面的复杂度取决于网络结构的设计和训练样本的分布；在支持向量机中，分类面的复杂度取决于核函数的选取及训练样本的分布。对于有限数目的训练样本来说，在决定神经网络结构和选择支持向量机核函数及其参数时，应充分考虑到分类器的推广能力问题，使模型和参数选择与样本和问题的复杂度相适应，必要时可以可以通过交叉验证等方法对可能的选择进行比较。

第六章讲述了除了前面两章介绍的线性和非线性分类方法的一些比较常用的方法，包括近邻法、决策树与随机森林、罗杰斯特回归和 Boosting 方法等。它们实现的分类器通常也是非线性分类器，但是他们的原理有一些特殊性。近邻法直接根据训练样本对新样本分类；决策树和随机森林可以应用于非数量特征，并把特征选择与分布决策结合起来；而罗杰斯特回归则是把分类问题转化成对概率的回归估计问题，Boosting 方法是通过将多个分类器进行融合从而得到更加有效的分类决策的方法。

第七章主要讲述了特征选择的特征评价准则，其中又介绍了基于类内类间距离的可分性判据、基于概率分布的可分性判据、基于熵的可分性判据和利用统计检验作为可分性判据。然后又介绍特征选择的最优算法、次优算法和遗传算法的原理方法，最后又介绍了以分类性能为准则的特征选择方法。如果特征之间存在冗余，通常在特征选择中会希望去掉这样的冗余，因为冗余的特征不能提供更多的分类信息。但是，在某些情况下，比如在利用基因表达数据进

行疾分类时，目的不仅仅是构造分类器把两类分开，而且希望通过这种分类鉴别出哪些基因的表达与所研究的疾病分类有关，也就是发现哪些特征与分类有关，此时就需要把所有包含分类信息的特征都保留下来进行下一步的分析，即使某种特征之间存在冗余。

第八章重点讲述了基于类别可分性判据的特征提取。然后介绍了主成分分析方法的原理和 K-L 变换的基本原理和应用方法。最后又介绍了多维尺度法的方法原理和应用，还介绍了非线性变换方法中的核主成分分析和 IsoMap 方法和 LLE 方法的原理。

第九章主要讲述了非监督模式识别的基本思想和一些有代表性的方法。重点讲述了混合模型的估计、动态聚类算法、模糊聚类方法、分级聚类方法和自组织映射神经网络。与监督模式识别相比，非监督模式识别问题中存在更大的不确定性，但同时也意味着非监督模式识别方法在人们探索未知世界中能够发挥更大的作用。在实际应用中，不但需要有效合理的运用本章介绍的非监督学习方法和其他聚类方法，而且要注意分析数据的特点，设法有效利用领域的专门知识，以弥补数据标注的不足。

第十章主要讲述了对模式识别系统的评价，其中重点讲述了监督模式识别方法的错误率估计、有限样本下错误率的区间估计问题等。评价模式识别系统性能的方法有多种，每种方法都有各自的特点和适用范围，当人们学习一种模式识别方法或者自己提出一种新的方法时，需要认真，客观的对待评价指标，选择能够反映问题本质的测试方法。