# 模式识别

中国科学技术大学 汪增福

- 第一章 绪论
- 第二章 统计模式识别中的几何方法
- 第三章 统计模式识别中的概率方法
- 第四章 分类器的错误率
- 第五章 统计模式识别中的聚类方法
- 第六章 结构模式识别中的句法方法
- 第七章 总结

# 统计模式识别小结与延伸


监督学习

非监督学习

半监督学习

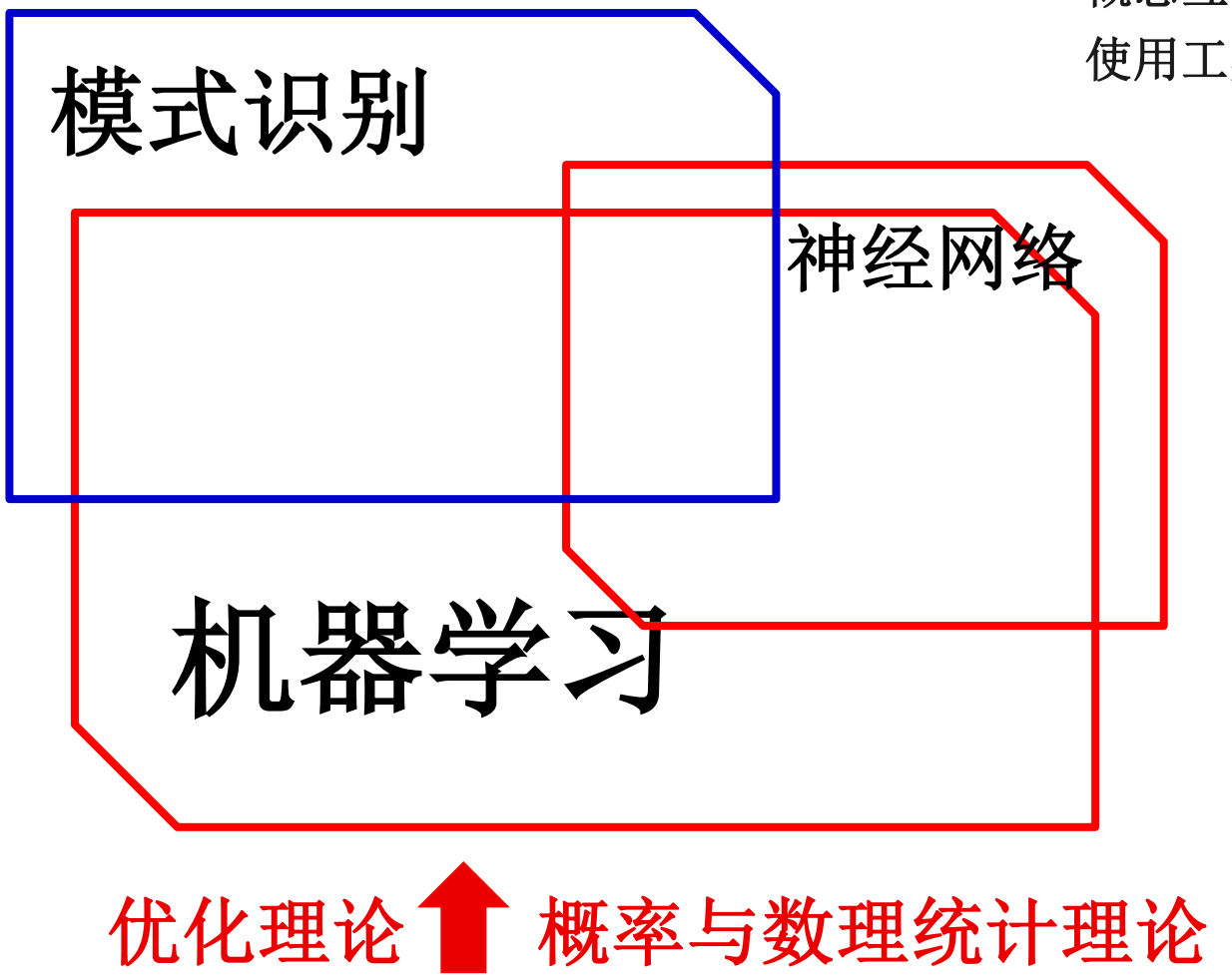
强化学习

神经网络

深度学习

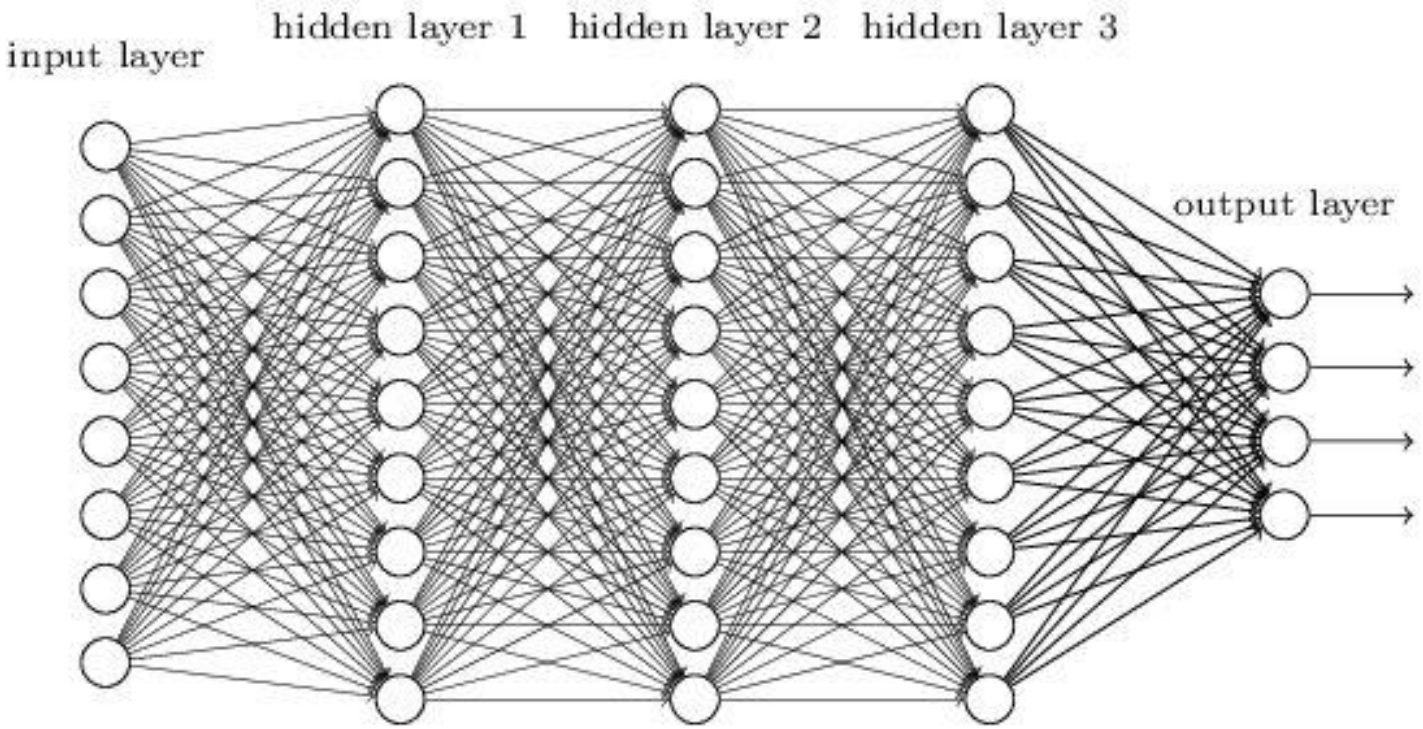
# 统计模式识别小结与延伸

产生背景不同
概念互有交叠
使用工具相通



# 统计模式识别小结与延伸

## 神经网络举例



# 统计模式识别小结与延伸

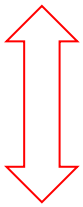
## 监督学习

用答案已知的例子（即标记过的数据）来训练神经网络，构建可用的分类器。

### Step 1 数据的生成和分组

把数据分为两组。第一组叫训练集，用来训练神经网络。第二组叫验证集，用来检验所构建分类器的性能。

### Step 2 训练神经网络



将训练集的数据输入（结构已知的）神经网络，并根据一定的规则决定某个神经元的输出，进而得到某一层的输出。当计算完所有神经元的时候，最后得到最右边的神经元的输出。

### Step 3 验证分类器

用验证集的数据验证训练得到的分类器的准确率。

# 统计模式识别小结与延伸

## 监督学习

- 分类
  - 从一组数据出发，确定输入样本和输出标签之间的对应关系。
  - 利用所建立的对应关系对输入样本的标签值进行预测。
- 回归
  - 从一组数据出发，确定输入变量和输出变量之间的定量关系式，建立所需要的数学模型并估计相关未知参数。
  - 利用所建立的数学模型预测输入样本对应的输出值。

# 统计模式识别小结与延伸

## 监督学习举例

- 从相册中找出包含你儿子的照片

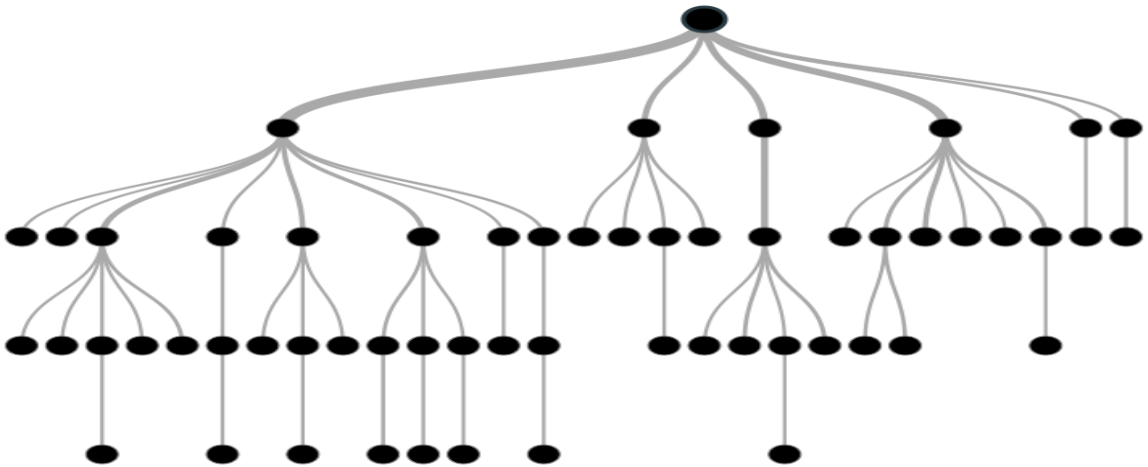
- 照片采集
- 照片标注
- 照片分组
- 分类器构建
- 分类器训练
- 分类器性能评估
- 应用分类器对收集到的所有照片进行整理

# 统计模式识别小结与延伸

## 监督学习之决策树方法

- 基于树的学习算法是最优秀，也是使用最广泛的监督型学习方法。基于树的算法赋予预测模型高精度、稳定性和易解释性。与线性模型不同，它们能很好地映射非线性关系，善于解决手头上的任何问题，包括分类和回归。

- 决策树是在分类问题中经常使用到的一种监督型学习算法。它适用于分类型以及连续性输入和输出变量。在这个方法中，我们根据输入变量中最显著的分裂点将总体或样本划分为两类或多类。



# 统计模式识别小结与延伸

## 监督学习之决策树方法

- 通过构造决策树来发现数据中蕴涵的分类规则。如何构造精度高、规模小的决策树是决策树算法的核心内容。

Step 1 决策树生成：由训练样本集生成决策树。

Step 2 决策树剪枝：对Step 1生成的决策树进行检验、校正和修剪。

用验证数据集中的数据校验决策树生成过程中产生的初步规则，将那些影响预测准确性的分枝剪除。

目标：根据给定的训练数据集构建一个决策树模型，使它能够对实例进行正确的分类。决策树学习本质上是从训练数据集中归纳出一组分类规则。能对训练数据进行正确分类的决策树可能有多个，可能没有。在选择决策树时，应选择一个与训练数据矛盾较小的决策树，同时具有很好的泛化能力；而且选择的条件概率模型应该不仅对训练数据有很好的拟合，而且对未知数据有很好的预测。

# 统计模式识别小结与延伸

## 决策树的基本思想

- 树以代表训练样本集的单个结点开始。
- 如果样本都在同一个类．则该结点成为树叶，并用该类标记。
- 否则，算法选择最有分类能力的属性作为决策树的当前结点。
- 根据当前决策结点属性取值的不同，将训练样本数据集分为若干子集，每个取值形成一个分枝，有几个取值形成几个分枝。针对上一步得到的一个子集，重复进行先前步骤，递归形成每个划分样本上的决策树。一旦一个属性出现在一个结点上，就不必在该结点的任何后代再考虑它。
- 递归划分步骤仅当下列条件之一成立时停止：

①给定结点的所有样本属于同一类。

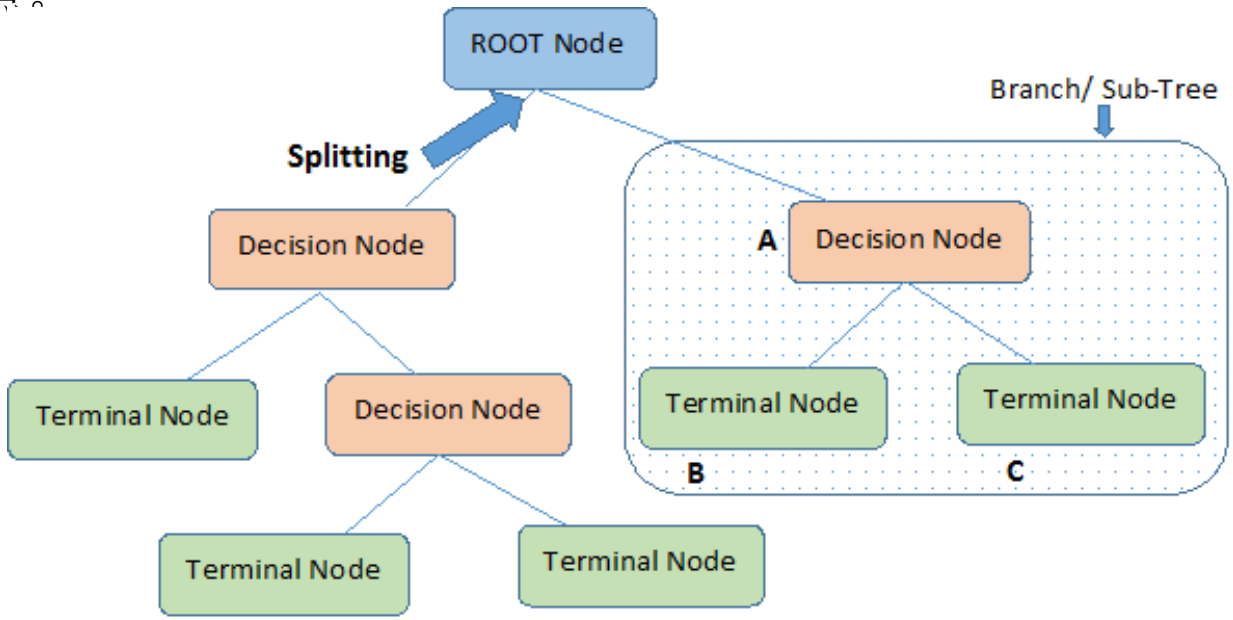
②没有剩余属性可以用来进一步划分样本。此时使用多数表决，将给定的结点转换成树叶，并以样本中元组个数最多的类别作为类别标记。

③如果某一分枝，没有满足该分支中已有分类的样本，则以样本的多数类创建一个树叶。

# 统计模式识别小结与延伸

## 决策树相关术语

- 根节点（Root Node）：代表整个总体或样本集。
- 分裂（Splitting）：将一个节点分成两个或多个子节点的过程。
- 剪枝（Pruning）：移除一个决策节点的子节点的过程。
- 决策节点（Decision Node）：当一个子节点（sub-node）进一步分成子节点时，我们称这个节点为决策节点。
- 叶子/终端节点（Leaf/Terminal Node）：不再分裂的节点称为叶子或终端节点。



# 统计模式识别小结与延伸

## 监督学习之决策树方法

- 决策树算法的优点

  分类精度高

  生成的模式简单，易于理解

  对噪声数据有很好的鲁棒性

  在数据探索过程中很有用：决策树是识别最显著变量以及两个或多个变量之间关系的最快速的方法之一。

- 决策树算法的缺点

  过拟合：过拟合是决策树模型中最为困难的问题之一。

# 统计模式识别小结与延伸

## 监督学习之随机森林方法

鉴于决策树容易过拟合的缺点，随机森林采用多个决策树的投票机制来改善决策树。

- 随机森林的生成方法
  - ✓ 从样本集中通过重采样的方式产生n个样本。
  - ✓ 假设样本特征数目为a，对n个样本选择a个中的k个特征，用建立决策树的方式获得最佳分裂点。
  - ✓ 重复m次，产生m棵决策树。
  - ✓ 对于分类问题，按多棵树分类器投票决定最终分类结果；对于回归问题，由多棵树预测值的均值决定最终预测结果。
- 随机森林法的优点：对于多维特征的数据集分类有很高的效率，并且可以做特征重要性的选择。运行效率和准确率较高，实现起来也比较简单。

# 统计模式识别小结与延伸

## 无监督学习

使用的所有数据均是没有进行标记的，即不知道输入数据对应的输出结果是什么。无监督学习只能默默的读取数据，自己寻找数据的模型和规律。

- 一个故事：Google Brain 团队的人想知道 YouTube 视频里面有没有猫。于是他们和斯坦福大学的 Quoc Le 和吴恩达合作，开发了一个聚类算法，能够把 YouTube 视频分为很多不同的类。他们并没有为系统设置“找猫”的任务，也没有提供任何标记数据，但是他们的算法能对 YouTube 视频自动分组，并且找出了猫。当然还有 ImageNet 所定义的22000个类别中的数千个其他物体。

- 假设你要生产T恤，却不知道 XS、S、M、L 和 XL 的尺寸到底应该设计多大。此时，你可以根据人们的体测数据，用聚类算法把人们分到不同的组，从而决定尺码的大小。

- 音乐鉴赏。

# 统计模式识别小结与延伸

## 半监督学习

半监督学习训练中使用的数据，只有一小部分是标记过的，而大部分是没有标记的。因此和监督学习相比，半监督学习的成本较低，但是又能达到较高的准确度。

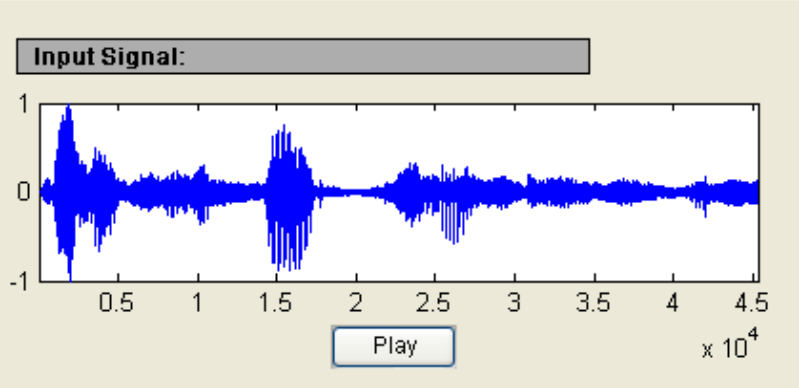
- Delip Rao用半监督学习的方法设计分类器。他对每个类别仅使用了30个有标记的数据，便取得了和用监督学习的方法对每个类别标记1360个数据一样的效果。

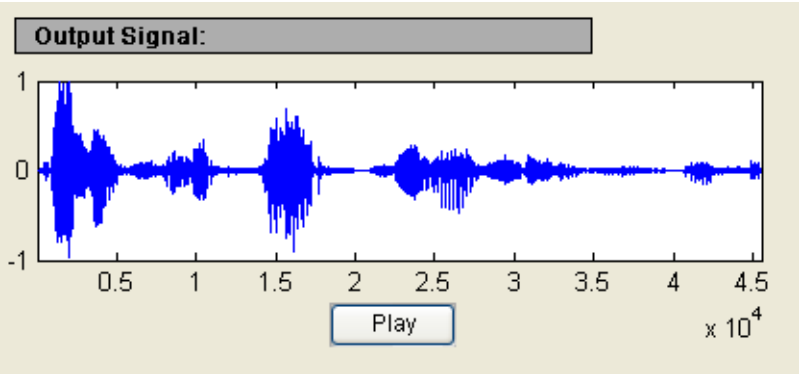
- 为什么无标记也能提高准确率的一个直观解释：即使不知道正确的答案，但是可以知道输入数据长什么样，有什么可能的取值。

# 统计模式识别小结与延伸

## 半监督学习

- **盲源分离**



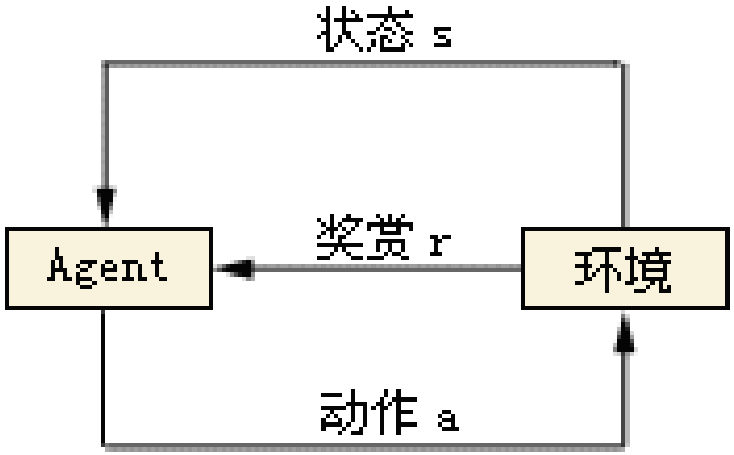
# 统计模式识别小结与延伸

## 强化学习

又称再励学习、评价学习，是一种重要的机器学习方法。

- **基本模型与原理**

  如果智能体的某个行为策略导致环境正的奖赏(强化信号)，那么智能体以后产生这个行为策略的趋势便会加强。智能体的目标是在每个离散状态发现最优策略以使期望的折扣奖赏和最大。

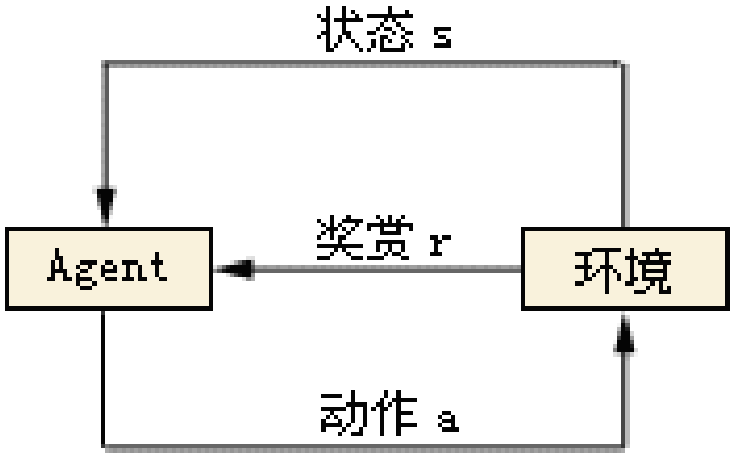


# 统计模式识别小结与延伸

## 强化学习

- **基本模型与原理**

由环境提供的强化信号不是告诉智能体如何去产生正确的动作。




强化学习把学习看作试探评价过程，智能体选择一个动作用于环境，环境接受该动作后状态发生变化，同时产生一个强化信号(奖励或惩罚)反馈给智能体，智能体根据强化信号和环境当前状态再选择下一个动作使受到正强化(奖励)的概率增大。

# 统计模式识别小结与延伸

## 强化学习

- **基本模型与原理**

  强化学习中，训练一段时间后，你才能得到一个延迟的反馈，并且只是提示说明你离答案是越来越远还是越来越近。

- **目标**

  学习从环境状态到行为的映射，使得智能体选择的行为能够获得环境最大的奖赏，使得外部环境对学习系统在某种意义下的评价(或整个系统的运行性能)为最佳。

若干图片和视频材料取自网络，特此致谢。

# 谢谢聆听！


中国科学技术大学