

**数据挖掘与知识发现**

## 第五次作业

**姓 名 覃书农**

**学 号 19041211954**

**专 业 控制科学与工程**

# 作业内容

天气因素有温度、湿度和刮风等，通过给出数据，使用决策树算法学习分类，输出一个人是运动和不运动与天气之间的规则树。给出的数据如图1-1所示。

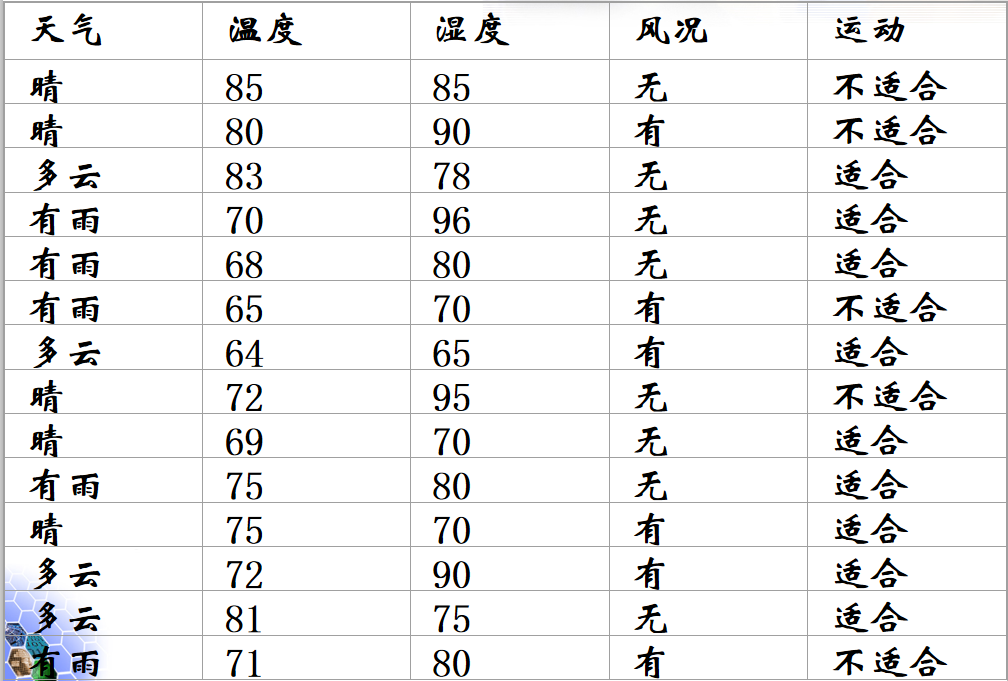


图1-1 天气数据

# 完成作业

## 2.1 归一化数据

因为决策树算法处理离散数据比较合适，因此要先将表中的温度和湿度两组数据归一化，在此约定：气温80及以上定义为“hot”，70~80定义为“mid”，70以下定义为“cool”，湿度90及以上定义为“high”，65~90定义为“normal”.

## 2.2 信息增益计算

根据公式2-1写出计算天气各个属性基于决策属性的期望的程序。

 2-1

根据式2-2、2-3以及式2-4写出选择最佳测试属性的程序。

 2-2

 2-3

 2-4

在选择最佳测试属性时，要注意避开已经选择过的属性。

## 2.3 特定属性抽取

为了便于统计和计算，需要基于特定属性的特定属性值进行抽取，例如，抽取数据中所有天气属性为晴的数据，进而可以统计天气为晴时是否适合运动的样本个数并计算期望和信息熵。

## 2.4 生成决策树

决策树通过多重字典的形式展示，采用函数自调用的方式实现字典的嵌套。结合上述算法构建决策树。

# 程序运行结果

程序输入的训练数据如下所示：

trainData = [  
 ['sunny', 85, 85, 'F', 'n'],  
 ['sunny', 80, 90, 'T', 'n'],  
 ['overcast', 83, 78, 'F', 'y'],  
 ['rainy', 70, 96, 'F', 'y'],  
 ['rainy', 68, 80, 'F', 'y'],  
 ['rainy', 65, 70, 'T', 'n'],  
 ['overcast', 64, 65, 'T', 'y'],  
 ['sunny', 72, 95, 'F', 'n'],  
 ['sunny', 69, 70, 'F', 'y'],  
 ['rainy', 75, 80, 'F', 'y'],  
 ['sunny', 75, 70, 'T', 'y'],  
 ['overcast', 72, 90, 'T', 'y'],  
 ['overcast', 81, 75, 'F', 'y'],  
 ['rainy', 71, 80, 'T', 'n']  
]

一共14组数据，每组数据从左到右分别为：天气（outlook）、温度（temperature）、湿度（humidity）、风况（windy）、是否适合运动。程序生成的多重字典为：

{'outlook': {'sunny': {'temperature': {'mid': {'humidity': {'normal': 'y', 'high': 'n'}}, 'cool': 'y', 'hot': 'n'}}, 'overcast': 'y', 'rainy': {'windy': {'T': 'n', 'F': 'y'}}}}

根据此字典，可以得到如图2-1的决策树。

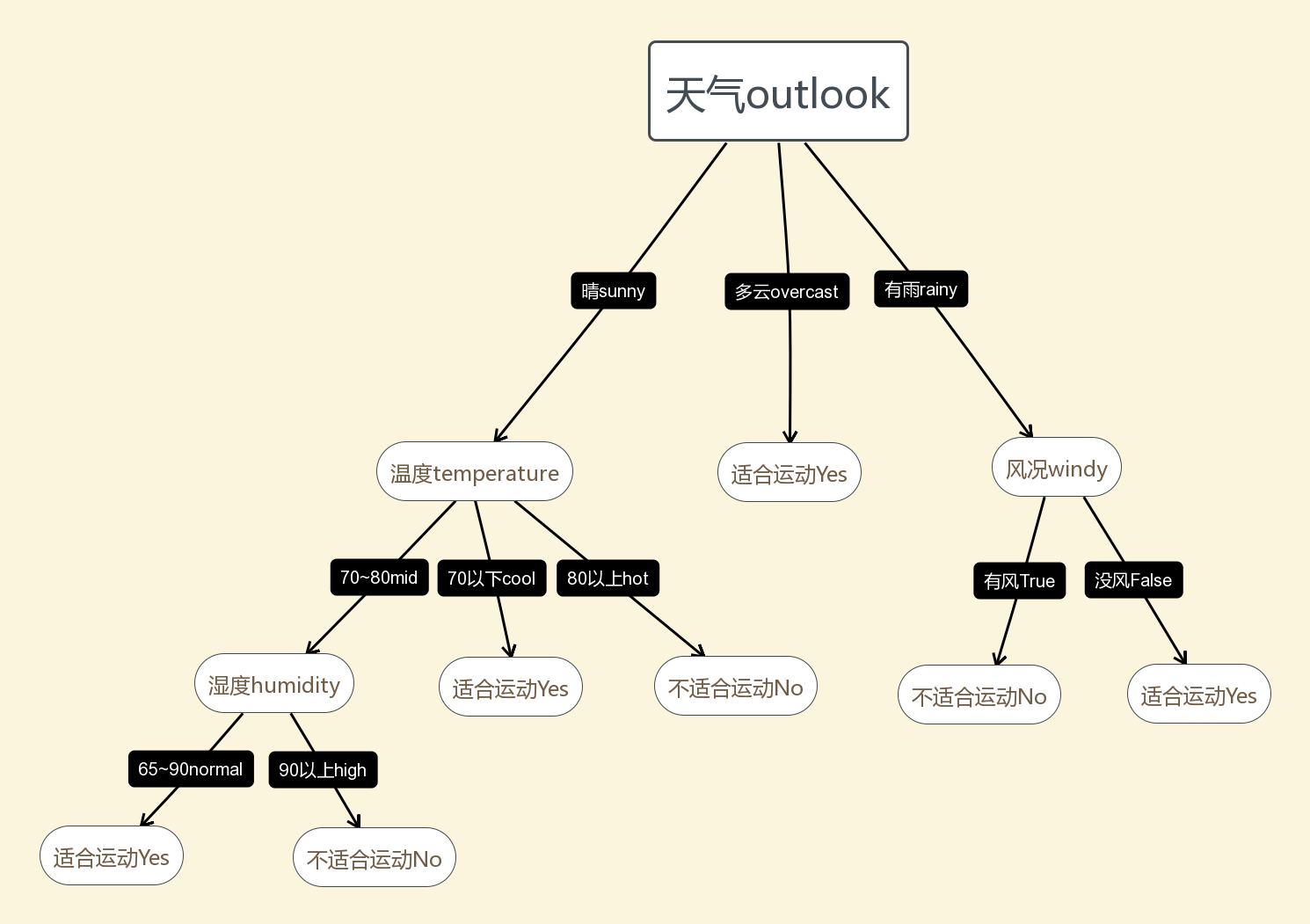


图2-1 决策树