# 大数据导论知识点汇总

——人工智能方法分析大数据

# 聚类分析

一、定义

-**簇**：数据对象的集合，同一簇中的对象之间彼此相似，不同簇之间的对象相异。

-**聚类分析**：把大型数据划分成不同的族。

-聚类是**无监督**分类：没有事先定义好的类别。

二、聚类的**应用场景**

-客户细分

-土地利用

-保险

-网络社区发现

三、好的**聚类方法**

-高的内部相似度（簇内越紧密越好）

-低的外部相似度（簇间越分离越好）

四、聚类质量**度量指标**

-Compactness（紧密性）

-Separation（间隔性）

-样本间相似性度量

五、主要的**聚类分析算法**

-划分法

-层次法

-基于密度的方法

-基于网格的方法

六、**K-means聚类算法**

①**步骤**

-从数据集中随机取k个对象作为k个簇的初始聚类中心。

-计算剩下的对象到k个簇中心的相似度，将这些对象分别划分到相似度最高的簇。

-根据聚类结果，更新k个簇的中心，计算方法是取簇中所有对象各自维度的算术平均值。

-将数据集中全部元素按照新的中心重新聚类。

-达到算法停止条件，转至步骤6；否侧转至3。

-输出聚类结果。

②算法**停止条件**

-设定迭代次数

-聚类中心不再变化

-前后两次聚类结果的目标函数变化很小

③**优劣**

-优点：效率较高

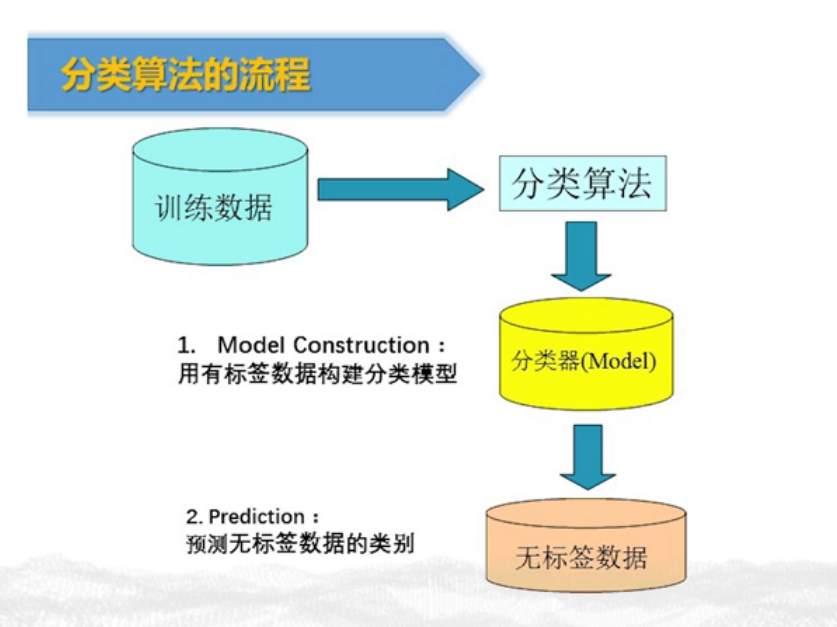
-不足：不适合于发现非凸形状的簇

# 分类分析

1. **分类算法的目标**

* 分类算法的目标是找到每个样本特征到类别的对应法则。​
* 前提是类别是已存在，即是有标签的数据，属于有监督学习类型。​
* 典型应用 : 信贷审批，目标市场，医疗诊断，欺诈检测等。​

1. **分类算法的流程**



1. **分类算法的训练和评价**

* 留出法(Holdout) :将数据集D划分为两个互斥的集合，其中一个集合作为训练集S,另一个作为测试集T。在S上训练出模型后，用T来评估其测试误差，作为对泛化误差的估计。
* 交叉验证(k-fold Cross-validation):将数据集分割成k个子样本。在每次运行时，使用一个不同的子样本作为测试集，其余的K-1子样本作为训练集。用k次运行的平均来估计这个方法的性能。
* 这种方法减少了训练集/测试集的随机性，有益于大数据集。

1. **主要分类算法**

* 决策树(Decision Tree)
* K邻近点分类(KNN)
* 贝叶斯分类(Bayes)
* 支持向量机(SVM)
* 人工神经网络(ANN)
* 逻辑回归(Logistic Regression)

1. **决策树分类算法**

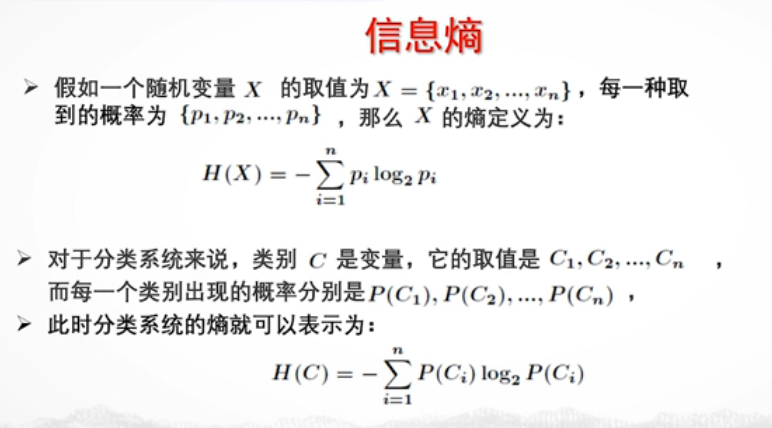
* 决策树通过训练数据构建决策树，对未知的数据进行分类。
* 决策树的每个内部节点表示在一个属性(Attribute)上的测试，每个分枝代表该测试的一个输出，而每个树叶结点存放着一一个类标号。
* 在决策树算法中，ID3基于信息增益作为属性选择的度量，C4.5基于信息增益比作为属性选择的度量，CART基于基尼指数作为属性选择的度量。

**六**

**1D3算法简介：**

* ID3算法是决策树的一种，即Iterative Dichotomiser3,迭代二叉树3代，是Ross Quinlan发明的一种决策树算法。
* ID3算法的核心思想就是以信息增益来度量属性的选择，选择分裂后信息增益最大的属性进行分裂。它使用信息熵(Entropy)定义的信息增益。
* 该算法采用自顶向下的贪婪搜索遍历可能的决策空间。

**信息熵：**



**信息熵的性质：**变量的不确定性越大，熵也就越大，把它搞清楚所需要的信息量也就越大。

**信息增益：**

* **信息增益**是针对一个属性而言的，待分类的集合的熵和选定某个属性的条件熵之差。在决策树分类问题中，信息增益就是决策树在进行属性选择划分前和划分后信息熵的差值。
* **划分规则：**

在决策树的每一个非叶子结点划分之前，先计算每一个属性所带来的信息增益，选择最大信息增益的属性来划分，因为信息增益越大，区分样本的能力就越强，越具有代表性。

**ID3算法的缺点：**

**·** **ID3的缺点：**信息增益偏向取值较多的属性

**·** **原因：**当特征的取值较多时，根据此特征划分更容易得到纯度更高的子集，因此划分之后的熵更低，因此信息增益更大，因此信息增益比较偏向取值较多的属性。

**·** **解决方法：**信息增益比( C4.5算法)



分裂信息Splitnfo用来衡量属性分裂数据的广度和均匀性，一个属性分割样本的广度越大，均匀性越强，该属性的分裂信息越大，增益率GainRatio就越小。因此分裂信息项降低了选择那些值较多且均匀分布的属性的可能性。

## 推荐系统

推荐系统——**Weka**

一、**简介**

Weka的全名是怀卡托智能分析环境，是一款免费的，非商业化的，基于JAVA环境下开源的机器学习以及数据挖掘软件。

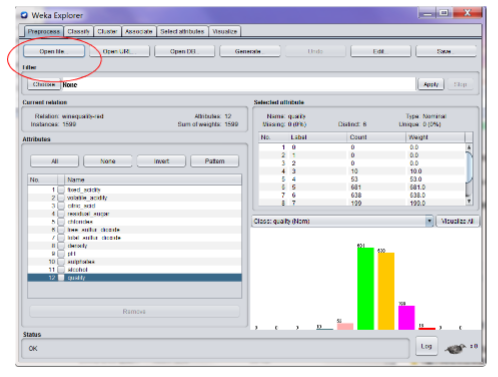
1. Weka下载网址

http://www.cs.waikato.ac.nz/ml/weka/downloading.html

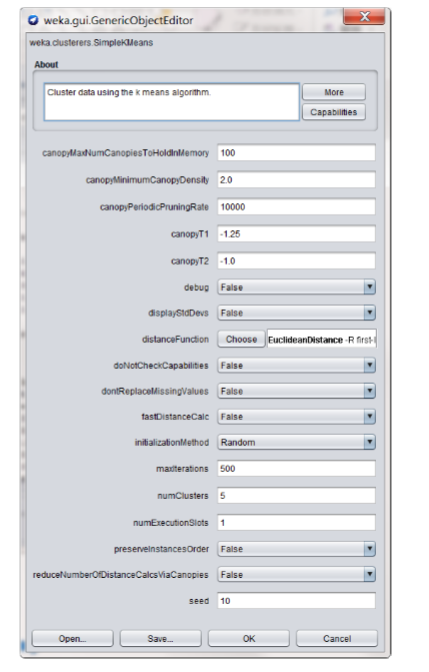
3、在聚类算法中的应用

①打开数据集

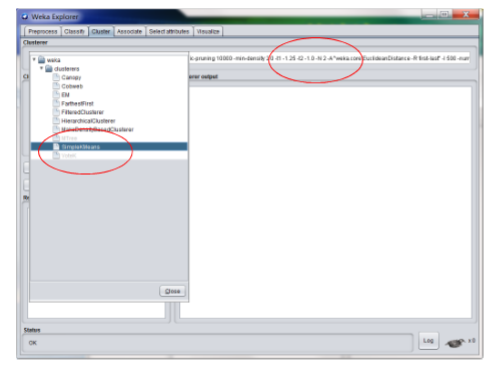
Weka所使用的数据集是一种以.arff为扩展名的文件



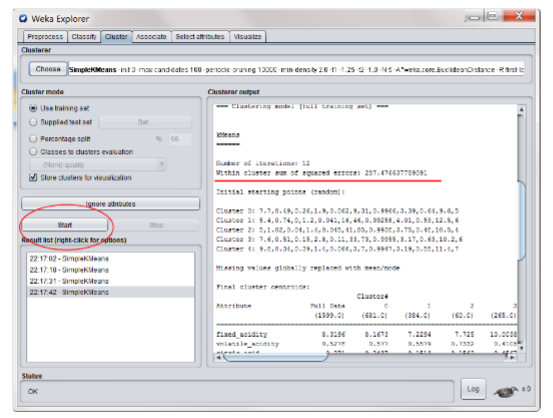
②设置算法参数



③选择算法—聚类



④呈现聚类结果



⑤聚类结果可视化

