链滴

数据挖掘算法初窥门庭--分类/回归

作者：Zing

原文链接：https://ld246.com/article/1457699791178

来源网站：链滴

数据挖掘算法中分类和算法经常放在一起，如weka中Classify包括了分类和回归。这两种方法都是通对已知类别训练集的分析，从中发现规律，以此预测新数据的类别。简单来说，若是预测的类别为离值则成为分类，若为连续值则成为回归。

分类/回归属于有监督学习，分为训练和预测两个过程（当然一般还会有模型的检验）。

- 训练：训练集->特征选取->训练->分类器模型
- 预测：新样本->特征选取->分类->判决

#决策树

决策树归纳是经典的分类算法。

决策树是将特征的判别序列形成一颗树，从树根到叶子节点进行每个节点的判断，叶子节点处对应某类别标号，就是最终的分类结果。

采用自顶向下递归的各个击破方式构造决策树。树的每一个结点上使用信息增益度量选择测试属性。以从生成的决策树中提取规则。

主要的决策树算法有：ID3,C4.5,CHAID,CART,Quest和C5.0

- 优点：
    - 模型易于理解和解释
    - 数据预处理阶段比较简单，可以处理缺失数据
    - 能够同时处理数值型和分类型数据
    - 能在相对短时间内对大数据集做出可行且效果良好的分类结果
- 缺点：
    - 对于那些各类别样本数量不一致的数据，在决策树当中,信息增益的结果偏向于那些具有更多数的特征。
    - 对噪声数据较为敏感
    - 容易出现过拟合问题
    - 忽略了数据集中属性之间的相关性

#KNN算法

KNN算法即K-最临近算法，是一种最简单实用的算法。

该方法的思路非常简单直观：如果一个样本在特征空间中的k 个最相似(即特征空间中最邻近)的样本的大多数属于某一个类别，则该样本也属于这个类别。

- 优点：
    - 简单有效，容易理解和实现
    - 重新训练的代价低
    - 计算时间和空间线性于训练集的规模
    - 适合处理多模分类和多标签分类问题
    - 对于类域的交叉或重叠较多的待分类样本集较为适合

- 缺点：
  - 是lazy学习方法，比一些积极学习的算法要慢
  - 对样本不平衡的数据集效果不佳，可以采用加权投票法改进
  - k值对分类效果影响很大，若K值太小会对噪声很敏感
  - 样本容量较小的类域采用这种算法比较容易产生误分

#SVM算法

支持向量机，可以自动寻找出那些对分类有较好区分能力的支持向量，并寻找一个超平面，最大化类类的间隔。

- 优点：
  - 对小样本的分类有较好的结果
  - 可以解决高维问题
  - 可以提高泛化性能
  - 可以解决非线性问题
  - 可以避免神经网络结构选择和局部极小点问题
- 缺点：
  - 对缺失数据敏感
  - 对非线性问题没有通用解决方案，必须谨慎选择核函数来处理

#贝叶斯分类

贝叶斯分类是利用贝叶斯公式，通过计算每个特征下分类的条件概率，来计算某个特征组合实例的分概率，选取最大概率的分类作为分类结果。

参见的贝叶斯分类器有：Naive Bayes, TAN, BAN, GBN等方法。

- 优点：
  - 基于完善的数学模型，分类效果稳定
  - 所需估计的参数很少，对缺失数据不太敏感
  - 无需复杂的迭代求解框架，适用于规模巨大的数据集
- 缺点：
  - 假设前提：属性之间独立性 往往不成立
  - 需要知道先验概率

#神经网络

神经网络是模拟人的神经反射功能，进行模型的自适应学习。通常分为输入层，输出层和中间层，通反馈对各层的参数进行调整和优化。

- 优点：
  - 分类准确性高，并行分布处理能力强
  - 对噪声有较强的鲁棒性和容错能力
  - 可以充分逼近非线性关系
  - 具备联想记忆能力
- 缺点：
  - 需要输入大量参数
  - 不能观察学习过程，输出结果难以解释
  - 学习时间长

#AdaBoost算法

提升方法是从弱学习算法出发，反复学习，得到一系列的弱分类器（即基本分类器），然后组合这些分类器，构成一个强分类器，大多数的提升方法都是改变训练数据集的概率分布（训练数据的权值分），针对不同的训练数据分布调用弱学习算法学习一系列的弱分类器。

- 优点：
  - 分类精度高
  - 可以使用各种方法构建子分类器
  - 简单，且不需要做特征筛选
  - 不会过拟合
- 缺点：
  - 对分类错误的样本多次被分错而多次加权后，导致权重过大，影响分类器的选择，造成退化问题
  - 数据不平衡问题导致分类精度的急剧下降
  - 算法训练耗时，拓展困难

#逻辑回归算法

二项logistic回归模型是一种分类模型，由条件概率分布P(Y|X)表示，形式为参数化的logistic分布。里随机变量X取值为实数，随机变量Y取值为1或0。可以通过有监督的方法来估计模型参数。

- 优点：
  - 计算代价不高
  - 易于理解和实现
  - 适用于数值型和分类型数据
- 缺点：

- 容易过拟合
  - 分类精度可能不高