# 实战项目 0_数据挖掘算法分析

**说明：本文系作者原创，他人不得抄袭！**

本章具体从算法原理分析算法局限性、实用性、适用数据类型、使用注意事项。

## §1.1 k-NN 近邻算法[7-9]

k-NN 算法是一种典型的基于实例的消极学习方法，以其简单有效和高鲁棒性而在分类问题中被广泛应用。

### 1.算法原理

对未知类别属性的数据集中的每个点依次执行以下操作：
（1）计算已知类别数据集中的所有点与当前点之间的距离；
（2）按照距离递增次序排序；
（3）选取与当前点距离最小的 k 个点；
（4）确定前 k 个点所在类别的出现频率；
（5）返回前 k 个点出现频率最高的类别作为当前点的预测分类。

距离计算公式为：

$$Dist\ (\mathbf{X},\mathbf{Y}) = \sqrt{\sum_{i=1}^{n}\left(X_i - Y_i\right)^2} \tag{1}$$

### 2.算法分析

**（1）局限性分析：**

局限 1：对于每一个待分实例，都要基于当前整个数据集计算待分实例与所有样本点间的距离；

局限 2：距离计算公式由平方、求和、开根号三部分运算组成，计算复杂度高；

局限 3：对于每一个待分实例，都要基于整个数据集进行排序，时间和内存开销大；

局限 4：近邻数目难以确定，对于非平衡数据集，容易产生较大的算法分类误差；

局限 5：对于高维数据，样本点间的距离趋于相同，因此不适用于高维数据；

局限 6：对于各维度数据范围差异大的数据，容易产生较大算法误差。

总结：计算复杂度高、空间复杂度高、k 值难以确定。

**（2）实用性分析：**

实用 1：算法采用“投票策略”，对异常值不敏感，算法整体精度高；

实用 2：算法无数据输入假定，适用性宽。

**（3）适用数据类型：**数值型和标称型

**（4）注意事项：**a.为了减小数据不同维度上数值范围的差异，应预先进行数值归一化或数据缩放等数据处理；

b.不适用高维数据和非平衡数据集。

## §1.2 朴素贝叶斯算法[10-13]

朴素贝叶斯基于“朴素假设”，计算特征概率，以极大似然估计思想选出最大概率的分类可能作为最终的分类结果，在文本分类中具有广泛的应用。

**1.算法原理**

**(1)算法基础：**2 个朴素假设

朴素假设（1）样本各维度的特征条件独立；

朴素假设（2）样本各维度特征对于分类作用相同。

**(2)概率计算转化：**设每个数据样本为：$\mathbf{X}=[X_1,X_2,...,X_n]$，$X_i$ 为第 i 个特征上的值，共 m 个类别，分别为$C_1$，$C_2$,...，$C_m$，对于待分实例$\mathbf{X}$，由贝叶斯定理得：

$$P(C_i \mid \mathbf{X})=\frac{P(\mathbf{X} \mid C_i)P(C_i)}{P(\mathbf{X})} \tag{2}$$

算法目标是最大化后验概率$P(C_i \mid \mathbf{X})$,因为$P(\mathbf{X})=\sum_{i=1}^{m}P(\mathbf{X} \mid C_i)P(C_i)$对于所有类为常数，所以，算法目标转化为最大化先验概率$P(\mathbf{X} \mid C_i)P(C_i)$。

**(3)计算先验概率：**基于“特征条件独立”这一朴素假设，简化计算$P(\mathbf{X} \mid C_i)$为：

$$P(\mathbf{X} \mid C)=\prod_{j=1}^{n}P(X_j \mid C_i)P(C_i) \tag{3}$$

**(4)极大似然估计：**待分实例所分类别为：

$$C_i=\arg\max_{i} c\, P(\mathbf{X} \mid C_i)P(C_i) \tag{4}$$

**2.算法分析**

**（1）局限性分析：**

局限 1：算法基于“特征对于分类作用平等”，忽略了不同类所关联的特征不同，而且类所关联的特征权重也不同的实际情况，算法准确率和适用性有限；

局限 2：算法基于“特征条件独立”，实际情况中，尤其是高维数据，特征间往往存在明显的关联性，所以，对于特征明显存在关联关系的数据并不适用；

局限 3：算法先验概率由特征条件概率乘积而得，对于非常小的概率，算法容易产生溢出；

局限 4：算法只能对待分实例给出最大可能分类结果，并不能给出每一个类别的关联特征和相应的权重值，算法作用单一。

局限 5：朴素贝叶斯在文本分类中应用广泛，但是，数据准备需要事先进行文本解析，所以，在某些应用场景中，数据准备方式对朴素贝叶斯影响也较大。

**（2）实用性分析：**

实用性 1：由于贝叶斯定理在实际情况中非常普遍，所以，对于较少的数据，朴素贝叶斯仍然有效，而且可以处理多类别问题；

实用性 2：算法简单高效，适用性广。

**(3)适用数据类型：**标称型数据

**（4）注意事项：**

a. 一定要对特征条件概率取对数，以防计算结果下溢；

b. 一定要对算法模型使用拉普拉斯平滑，以防算法对于未知类别样本分类错误；

c. 由于特征条件独立假设在实际应用中不满足，所以，实际应用中可以通过降维技术如 PCA 降维来降低朴素假设影响；

d. 由于特征平等假设在实际应用中不满足，所以，要么使用基于加权的朴素贝叶斯，要么同其他算法如决策树来事先确定特征权重。

e. 某些应用场景中，注意数据格式的准备，如文本分类中要注意文本解析。

## §1.3 C4.5 决策树算法[14-19]

决策树算法应用非常广泛，其树形结构直观便于理解，对于数据集的分布没有背景要求，算法准确率高，是一种实用性很高的监督分类算法。

**1.算法原理**

(1)检测数据集中的每个子项是否属于同一分类；

(2)If　so return 类标签

(3)EIif 遍历完数据集中所有特征

(4)基于“投票策略”返回频度最高的类标签

(5)Else

(6)寻找划分数据集的最好特征

(7)根据最好特征划分数据集

(8)创建分支节点

(9)For 每个划分的子集

(10)从（1）开始重新操作

C4.5 算法信息熵计算公式：

$$Ent(X) = -\sum_{i=1}^{v} P_i \log_2 P_i = -\sum_{i=1}^{v} \frac{|S_i|}{S} \log_2 \frac{|S_i|}{S} \tag{5}$$

式中，X 为特征，v 是特征 X 取值的总数，$P_i$ 为特征 X 取值为 i 时的样本数在当前总样本中所占频率，S 为当前样本总数，$|S_i|$ 为特征 X 取值为 i 时的样本数。

C4.5 算法分离信息即信息增益率分母的计算公式：

$$SplitInformation(S, X) = -\sum_{i=1}^{v} \frac{|S_i|}{S} \log_2 \frac{|S_i|}{S} \tag{6}$$

式中，v 是特征 X 取值的总数，$|S_i|$为特征 X 取值为 i 时的样本数，S 为当前样本总数。

**2.算法分析**

**（1）局限性分析：**

局限性（1）：对于特征过多的数据集，决策树的过拟合问题非常严重；

局限性（2）：决策树采用贪心策略，容易获得局部最优值而无法达到全局最优值；

局限性（3）：决策树一旦构造完毕，除非重新构造决策树，否则很难在现有决策树基础上更改树的结构；

局限性（4）：对于数值型数据，仍需离散化，但是，怎样离散化数值值获得最好效果仍然需要多次尝试；

局限性（5）：寻找划分数据集最佳特征时，决策树逐个试探，没有使用良好的启发式搜索策略，搜索效率低。

局限性（6）：C4.5 决策树需要保存训练好的树结构，所以，对于大容量数据集来说，C4.5 决策树可能并不“实用”。

**（2）实用性分析：**

实用性（1）：决策树容易写成 If-Else 结构，便于直观理解；

实用性（2）：引入惩罚项分离信息，避免倾向于选择有较多属性值的特征；

实用性（3）：可以使用决策树进行预处理，筛选出冗余特征和不相关特征；

实用性（4）：可以有效和智能搜索方法结合，获得更好的算法效果；

实用性（5）：对于小数据集，计算效率高、算法准确率高。

**（3）适用数据类型：**数值型数据、标称型数据

**（4）注意事项：**

a.数值型数据连续化，仍需多次尝试；

b.缺失值处理，可查阅相关文献，具体情况具体分析；

c.决策树剪枝的选择，可查阅相关文献，具体情况具体分析。

## §1.4 k-Means 算法[20-24]

k-Means 算法在无监督聚类中具有非常广泛的应用，在进行有监督学习任务前，往往需要使用 k-Means 等无监督学习方法来探索数据集的内在数据结构，基于属性的某种相似性度量标准，将样本划分到最优簇中。相似性度量是聚类算法的核心。实际应用中，往往需要结合具体的数据类型和数据挖掘需求，选择合适的聚类算法。所以，对于具体的应用问题，常常需要先试用几种不同的聚类算法，然后再基于聚类效果和算法时空复杂度来选出“最优”的聚类算法。

**1、算法原理**

**算法基础（隐含假设）：**kMeans 算法内在假设数据集是由 k 个球（或超球）分量混合而成的，每一个聚簇对应一个分量。

（1）创建 k 个点作为 k 个簇的起始中心（随机搜索或启发式搜索都可以）

（2）当任意一个点的簇分配结果发生改变时

（3）　　对数据集中的每个数据点

（4）　　对每个质心

（5）　　计算质心与数据点间的距离

（6）　　将数据点分配到距其最近的簇

（7）　　对每一个簇，计算簇中所有点的均值并将均值作为质心，跳转（2）

算法目标优化函数为：

$$\sum_{i=1}^{m}\sum_{j=1}^{k}(\arg\min_{j}\left\|X_i - C_j\right\|_2^2) \tag{7}$$

式中，j={1,2,...,k}，k 为聚类总数目，$X_i$ 为第 i 个样本，i={1,2,...,m}，m 为样本总数，$C_j$ 为第 j 个簇的质心，$\left\|X_i - C_j\right\|_2^2$ 为第 i 个样本与第 j 个簇质心的距离平方，目标函数是：使得所有样本到各自簇质心的距离平方和最小。

**2、算法分析**

**（1）局限性分析：**

局限性（1）：k-Means 算法本质上是一种面向非凸代价函数优化的贪婪下降求解算法，所以，仅能获得局部最优解；

局限性（2）：对初始聚簇中心非常敏感。同样的数据集，不同“质量”的初始簇中心，聚类质量也会相应有所不同；

局限性（3）：实际应用中，选择最优的 k 值比较困难，因为数据集的真实内在结构是隐含的。

局限性（4）：不同的数据集具有不同的内在结构特点，如何选择符合数据集分布特点的相似性度量标准仍然需要多次尝试；

局限性（5）：某些情况下，为了避免算法获得局部最优解，需要事先设立

算法停止准则。如何设立一个合适的停止准则，需要多次尝试；

局限性（6）：如果实际的数据集并不是若干球形高斯分布的重叠，则k-Means 算法不稳定；

局限性（7）：因为要最小化目标函数，所以，k-Means 算法使用均值统计量。又因为均值对异常值和噪声数据敏感，不是一种稳健的统计量，因此，k-Means 算法对异常值和噪声数据敏感；

局限性（8）：kMeans 算法可能产生“空聚簇”，尤其对于高维数据，数据分布非常稀疏，使得某一空间中会产生“空簇”。

局限性（9）：聚类有时会产生很小的“冗余簇”，实际情况中，要有选择的进行合并。怎么合并最适合，仍需要进一步探讨；

局限性（10）：算法仅能将数据集分成若干个簇，不能给出每个簇（类）相关联的特征及其相应的特征权重；

局限性（11）：算法忽略了不同特征和不同簇（类）之间的关联关系，同时也忽略了同一簇（类）内不同特征的重要度不同的实际情况；

局限性（12）：某些相似性度量标准如 L2 距离，容易受大值属性左右，所以，对于特定的相似性度量标准，应采取相应的数据预处理；

局限性（13）：对于高纬度数据，算法计算复杂度高，资源开销大。

**（2）实用性分析：**

实用性（1）：算法原理简单，通常情况下是有效的；

实用性（2）：算法可伸展性好，只要使用合适的相似性度量标准，就可以对于具有不同数据分布特点的数据集获得良好的聚类效果；

实用性（3）：算法应用性广，且算法易于实现。

**（3）适用数据类型：**数值型数据

**（4）注意事项：**

a.聚类数目 k 的确定，可以事先通过某种途径获得关于簇数目的先验知识，如事先观察、人为指定；也可以设定为 PCA 降维[25]后的主成分个数，主要考虑每一个主成分代表一个类；也可以使用半监督聚类技术[26-27]；也可以穷举，多次聚类选择最佳聚类结果，但是，要注意产生“冗余簇”的影响；也可以与其他算法相结合，如：在 k-Means 聚类结果上再进行层次聚类[28]；也可以修改聚类优化目标函数，使得算法自动获得最佳聚类数目；也可以基于模型选择“偏差-方差权衡”确定最佳聚类数目；

b.如果实际数据集不是多个球状簇的叠加形成，那么可以使用数据降维或数据缩放如“白化”缩小数据间的距离差异，使数据趋于球状分布的叠加效果；或者，

针对数据集的分布特点，选择合适的相似度度量标准，多次尝试选择最佳效果；

c.如果数据有大值属性，为了消除不同属性间的取值范围差异，需进行 0-1 标准化或数据缩放等数据预处理；

d.对于数据预处理，应包含：处理缺失值、去除异常点和冗余数据、降维。

e.以防算法获得局部最优值，可以基于不同的初始簇中心多次运行算法，选择最佳聚类效果；

f.如果当前数据集不可分，则可以选择合适的核函数[29-32]将数据集映射到高维空间中，使得其线性可分，但是，要注意：在高维空间中，谨慎使用距离度量标准，因为在高维空间中，不仅有可能产生“空簇”，还会发生各数据点间的的距离趋于相同的现象，使得聚类无法进行；而且，将数据映射到高维空间后，为了减小数据分布稀疏性的影响，往往需要增加样本数量，这也是要注意的；最后，核函数的选择非常灵活，怎样选择合适的核函数，也需要多次尝试才能获得；

g.初始聚类中心的选择，可以结合其他智能优化算法如遗传算法等，减小初始类中心对算法的收敛波动性影响；也可以利用半监督聚类技术；也可以事先考察，认为设定；也可以基于多组初始聚类中心，多次运行算法，选择最佳聚类效果；

h.对于大数据集，应为 k-Means 算法加速。具体请阅读相关专业文献。

## §1.5 本章小结

本章简要介绍了算法原理，并从算法局限性、算法实用性、适用数据类型、注意事项 4 个角度对算法进行了总结，具有实际意义。