模式识别

第一章

第一个分类利器

K最近邻(K-NEAREST NEIGHBOR)
KNN分类算法

naive, simple, graceful

主要内容

1 引言

2 KNN的基本思想

3 KNN算法的实现

4 KNN的优缺点

5 KNN的一些改进策略

6 KNN在实际问题中的应用

1 引言

分类（Classification）是一种重要的技术，它是从一组已知的训练样本中发现分类模型，并且使用这个分类模型来预测待分类样本。建立一个有效的分类算法模型最终将待分类的样本进行处理是非常有必要的。

目前常用的分类算法主要有：贝叶斯分类算法（Bayes）、支持向量机分类算法（Support Vector Machines）、 KNN最近邻算法(k-Nearest Neighbors)、神经网络算法（NNet）以及决策树（Decision Tree）等等。

KNN算法是一个理论上比较成熟的方法，最初由Cover和Hart于1968年提出，其思路非常简单直观，易于快速实现。

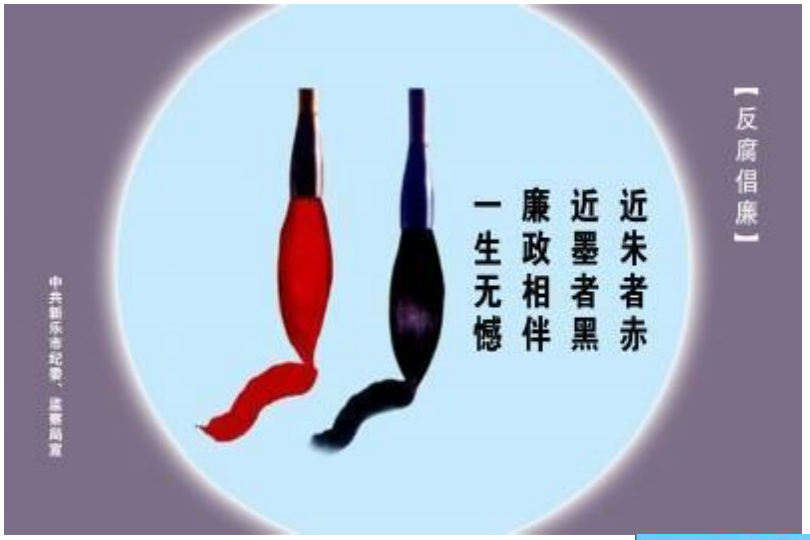
因此，KNN算法以其实现的简单性及较高的分类准确性在中文文本自动分类等领域得到了广泛应用。

1.1 最近邻法

- 以全部训练样本作为“代表点”，计算测试样本与这些“代表点”，即所有样本的距离，并以最近邻者的类别作为决策。这种方法就是近邻法的基本思想。
- 将与测试样本最近邻样本的类别作为决策的方法称为最近邻法。

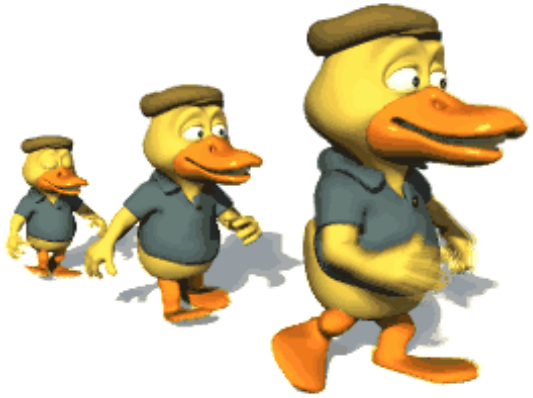
最近邻法

- 思想：“近朱者赤近墨者黑”



最近邻法

- 合理性：*相似的类别具有相似的特别*



最近邻法

- “理论基础”：

训练样本足够多的情况下，测试样本与其最相近的训练样本应具有相同的类别

1.1 最近邻法

- **最近邻法：**

- 将测试样本分类为与其距离最近的训练样本的类别

1.1 最近邻法

- **最近邻法的决策规则如下**

$$g_i(x) = \min_k \| x - x_i^k \|, k = 1, 2, \cdots, N_i$$

$$\text{如果} g_j(x) = \min_i g_i(x) \quad \text{那么：} x \in \omega_j$$

1.1 最近邻法

- 计算错误率的偶然性会随着训练样本数量的增大而减少，就利用训练样本数量增至极大，来对其性能进行评价。也就是在渐进概念下分析错误率。

- 理论上，最近邻法的错误率小于最小错误分类(贝叶斯分类)错误率的两倍。

1.1 最近邻法

- 不足：
- 最近邻法存在计算量大，存储量大等明显缺点。
- 训练样本集的数量总是有限的，有时候多一个或者少一个训练样本将会对测试样本分类的结果产生较大的影响。

2 KNN的基本思想

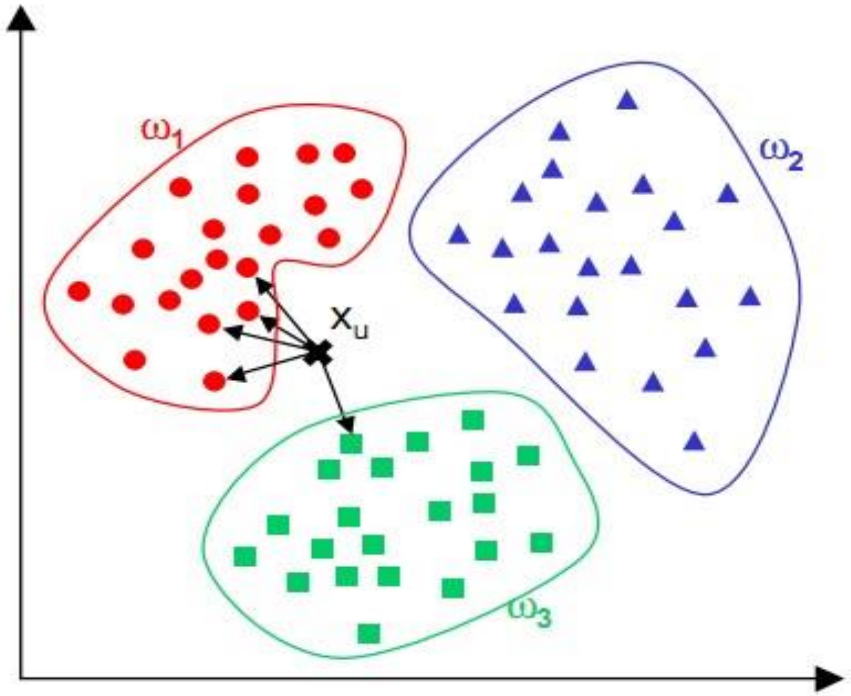
根据距离函数计算待分类样本X和每个训练样本的距离（作为相似度），选择与待分类样本距离最小的K个样本作为X的K个最邻近，最后以X的K个最邻近中的大多数所属的类别作为X的类别。

KNN可以说是一种最直接的用来分类未知数据的方法。

思想：少数服从多数

更多解释：

简单来说，KNN可以看成：有那么一堆你已经知道分类的数据，然后当一个新数据进入的时候，就开始跟训练数据里的每个点求距离，然后挑出离这个数据最近的K个点，看看这K个点属于什么类型，然后用少数服从多数的原则，给新数据归类。


ω1
ω2
ω3
Xu

3 KNN算法的实现

(1)问题描述

数据集：iris.data标准数据集-鸢尾花。

采用KNN算法对iris.data分类。为了操作方便，对各组数据添加rowNo属性，第一组rowNo=1，共有150组数据,选择rowNo模3不等于0的100组作为训练数据集，剩下的50组做测试数据集。

(2)实现步骤：

①初始化距离为最大值；

②计算未知样本和每个训练样本的距离dist；

③得到目前K个最临近样本中的最大距离maxdist；

④如果dist小于maxdist，则将该训练样本作为K-最近邻样本；
⑤重复步骤2、3、4，直到所有未知样本和所有训练样本的距离都算完；
⑥统计K-最近邻样本中每个类标号出现的次数；
⑦选择出现频率最大的类标号作为未知样本的类标号。

4 KNN的优缺点

- 优点

(1) 算法思路较为简单，易于实现；

(2) 当有新样本要加入训练集中时，无需重新训练（即重新训练的代价低）；

(3) 计算时间和空间线性于训练集的规模（在一些场合不算太大）。

◆ 不足

(1)分类速度慢；

KNN算法的时间复杂度和存储空间会随着训练集规模和特征维数的增大而快速增加。因为每次新的待分样本都必须与所有训练集一同计算比较相似度，以便取出靠前的K个已分类样本。整个算法的时间复杂度可以用O(m*n)表示，其中m是选出的特征项(属性)的个数，而n是训练集样本的个数。

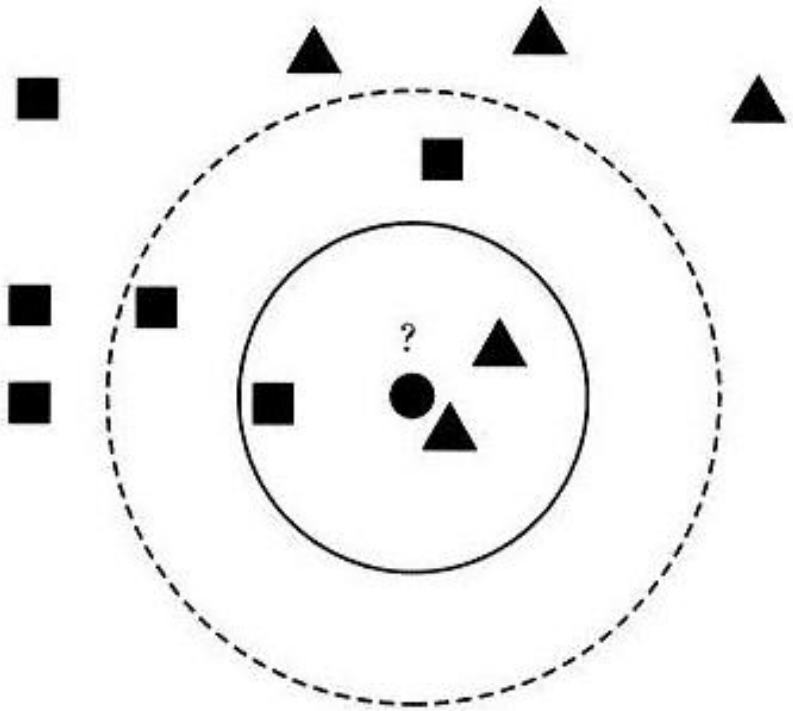
(2) 各属性的权重相同，影响了准确率；

(3) 当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。该算法只计算“最近的”邻居样本，如果某一类的样本数量很大，那么可能目标样本并不接近这类样本，却会将目标样本分到该类下，影响分类准确率。

(3) 样本库容量依赖性较强；

(4) K值不好确定；

k值选择过小，得到的近邻数过少，会降低分类精度，同时也会放大噪声数据的干扰；而k值选择过大，如果待分类样本属于训练集中包含数据较少的类，那么在选择k个近邻的时候，实际上并不相似的数据也被包含进来，造成噪声增加而导致分类效果的降低。



5 KNN的一些改进策略

(1)从降低计算复杂度的角度

当样本容量较大以及特征属性较多时，KNN算法分类的效率就将大大降低。可以采用以下方法进行改进。

✓如果在使用KNN算法之前对样本的属性进行约简，删除那些对分类结果影响较小（不重要）的属性，则可以用KNN算法快速地得出待分类样本的类别，从而可以得到更好的效果。

- ✓ 缩小训练样本的方法：在原有的样本中删掉一部分与分类相关不大的样本，将剩下的样本作为新的训练样本或者在原来的训练样本集中选取一些代表样本作为新的训练样本；
- ✓ 通过聚类（clustering），将聚类所产生的中心点作为新的训练样本。

(2)从优化相似度度量方法的角度

基本的KNN算法基于欧几里得距离来计算样本的相似度，这种方法对噪声特征非常敏感。

为了改变传统KNN算法中特征作用相同的缺陷，可在度量相似度的距离公式中给特征赋予不同权重，特征的权重一般根据各个特征在分类中的作用设定。

(3)从优化判决策略的角度

传统的KNN算法的决策规则的缺点是，当样本分布不均匀（训练样本各类别之间数目不均衡，或者即使基本数目接近，由于其所占区域大小的不同）时，只按照前K个邻近顺序而不考虑它们的距离，会造成误判，影响分类的性能。

可以采用均匀化样本分布密度的方法进行改进。

(4) 从选取恰当k值的角度

由于KNN算法中几乎所有的计算都发生在分类阶段，而且分类效果很大程度上依赖于k值的选取。而目前为止，比较好的选k值的方法只能是通过反复试验调整。

k值增大到一定程度后，会带来分类正确率的下降。

小结：

KNN算法简单，易于实现，但当样本规模很大时，其复杂度会很大，所谓“适合的就是最好的”，在选择分类算法时我们应该根据具体应用的需求，选择适当的分类算法。

作业与思考：

1. 你对“过适应”与“泛华能力”差的了解？

2. 模型和算法训练过程中有哪些手段可以避免或减轻过适应？

3. K近邻分类方法中K的取值应该注意什么问题？

4. 字符识别数据集最近邻分类的实现， 并通过实验说明。训练集大小对结果的影响。

5. 字符识别数据集K近邻分类的实现，并通过实验分析不同K值的影响.

6. 尝试对传统的K近邻分类进行改进，并给出实验结果对比。