K 近邻算法实验报告

张峻伟
（重庆大学软件学院， 重庆， 401331）

摘要：KNN 算法是一种基本分类与回归方法，是著名的模式识别统计方法，是最好的文本分类算法之一，在机器学习分类算法中占有相当大的地位。本文对 KNN 算法做了一份实验，详细介绍 KNN 算法的思想、原理、实现步骤以及具体实现代码，并分析了算法的优缺点。
关键词：K 近邻算法；KNN 算法；机器学习；统计学习方法

1 简介

分类是数据挖掘中、自然语言处理等多个领域的核心和基础技术，在经营、决策、管理、科学研究等多个领域都有着广泛的应用。目前主要的分类技术包括决策树、贝叶斯分类、KNN 分类等。在这些方法中，KNN 分类是一种简单、有效、非参数的方法，现已经广泛应用于文本分类、模式识别、图像及空间分类等领域。本文从各个角度对 KNN 算法进行较为全面的总结。

KNN 算法的指导思想是“近朱者赤，近墨者黑”，由数据周围的邻居来推断出数据的类别。这其中主要有三个基本要素，分别是：距离度量、K 值、分类决策规则，确定了三个基本要素也就确定了数据的分类类别。

2 算法方法

2.1 算法计算步骤

1. 算距离：给定测试对象，计算与训练集中的每个对象的距离；
2. 找邻居：圈定距离最近的 K 个训练对象，作为测试对象的近邻；
3. 做分类：根据这 K 个近邻归属的主要类别，来对测试对象分类。

2.2 距离度量

1. 欧式距离：

$$\mathrm{d}_{\mathrm{euc}}\ x,y\ = [\sum_{j=1}^{d}(x_j - y_j)^2]^{\frac{1}{2}} = [\ x-y\ (x-y)^T]^{\frac{1}{2}} \quad (1)$$

3. 曼哈顿距离

$$d = |x_1 - x_2| + |y_1 - y_2| \quad (2)$$

4. 闵式距离

$$d(x,y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

5. Lp 距离：

$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^p \right)^{\frac{1}{p}} \quad (4)$$

2.3 K 值的选择

与实例相似的实例个数为 k，当 k 较小时，近似误差减小，估计误差增大，易受噪声污染和过拟合；一般采用小的 K 值，再采用交叉验证法。

2.4 分类决策规则

投票决定：少数服从多数，近邻中哪个类别的点最多就分为该类加权投票法：根据距离的远近，对近邻的投票进行加权，距离越近则权重越大（权重为距离平方的倒数）

2.5 算法优缺点

算法优点为简单，易于理解，易于实现，无需估计参数，无需训练，适合对稀有事件进行分类；特别适合于多分类问题。但是作为懒惰算法，对测试样本分类时的计算量大，内存开销大，评分慢；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

2.6 常见问题

1. K 值的选择

K 值的选择过小，得到的近邻数也越少，会降低分类精度，同时也会放大噪

声数据的干扰，而如果 K 值选择过大，并且待分类样本属于训练集中包含数据较少的类，那么在选在 K 个近邻的时候，实际上并不相似的数据亦包含进来，造成噪声增加而导致分类效果的降低。

2. 类别的判定方式

投票法没有考虑近邻的距离的远近，距离更近的近邻也许更应该决定最终的分类，所以加权投票法更恰当一点。

3. 距离度量方式的选择。

高维度对距离衡量的影响：众所周知当变量数越多，欧式距离的区分能力就越差，所以值域越大的变量常常会在距离计算中占据主导作用，因此应先对变量进行归一化处理。

3 实验

在这部分内容中，我们使用鸢尾花数据集，选择欧式距离度量，采用投票表决规则介绍 KNN 算法的算法流程，实现过程以及测试数据效果的展示，通过设置不同的 K 值，对测试数据集的准确率等性能参数进行比较，验证 K 值的选择对 KNN 算法的影响，下面对实验细节和实验结果进行详细描述。

3.1 算法流程

KNN 算法解决鸢尾花分类问题

输入：训练数据集 X，测试数据集 Y，设定 K 值

输出：测试数据集的标签类型 answer，召回率，准确率以及 F1

初始化：对训练数据集中的数据进行归一化处理，将标签字符串转化为数字，对测试数据集中的数据进行归一化处理。计算每类训练数据集与测试数据集中数据的距离，并进行排序。

重复：

步骤一：计算 K 个样本所属的类别个数

步骤二：选择出现的类别次数最多的标签

直到遍历完数据内的 K 个数据

计算召回率，准确率，精确率以及 F1

3.2 实验数据与设置

鸢尾花数据集是机器学习中使用最普遍的数据集之一，它包含了花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。我们选取了 104 项数据形成训练集，46 项数据形成测试集，通过对比测试数据集中数据与训练数据集中每个数据中 4 个属性的欧式距离进行分类，选择前 K 个欧式距离最小的数据，在 K 个数据中根据投票表决规则选择测试数据集中数据的类别。

图 1. 展示了训练集中鸢尾花种类的分布。

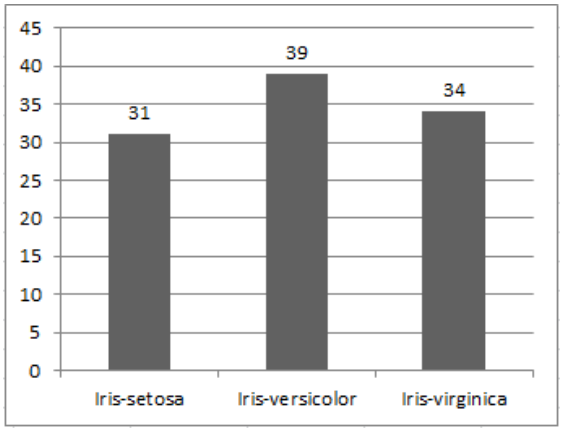
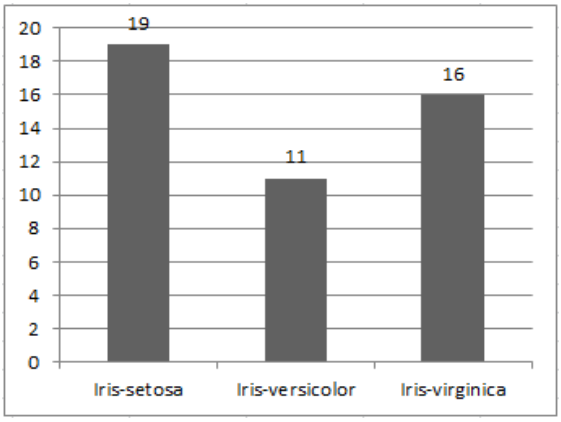


图 2. 展示了测试集中鸢尾花种类的分布。



为了验证不同 K 值的选择对 KNN 算法的影响，我们分别设置不同 K 值进行对比，设置类别为 Iris-virginica 的类为正类，其他两种类别为负类。从准确率、精确率、召回率和 F1 值可以看出，K 值的选择不同，随着 K 值选择的增大，预测准确率和精确率应该逐步增加，但是当 K 为 40 时，准确率和精确率降低，召回率降低，出现了将正类预测为负类的情况。选择较大的 K 值，就相当于用较大邻域中的训练实例进行预测，本应该可以减少学习的估计误差，但是由于与测试数据距离较远的训练数据也对预测起到了作用，使得预测发生错误，准确率反而降低。所以，K 值的选择是影响 KNN 算法分类性能的重要因素。

表 1. 展示了不同 K 值， KNN 分类算法的准确率、精确率、召回率和 F1 值

K 值	准确率	精确率 P	召回率 R	F1 值
1	93.478	0.842	1	0.914
5	95.652	0.889	1	0.941
10	95.652	0.889	1	0.941
20	97.826	0.941	1	0.970
40	95.652	0.938	0.938	0.938

4 总结

在这篇实验报告中，我们对 KNN 算法的三个基本要素和常见问题进行了总结，采用鸢尾花数据对 KNN 算法进行了分析与实践，通过选择不同的 K 值，观察 KNN 算法的准确率、精确率、召回率和 F1 值的变化，从而得出 K 值的选择是影响 KNN 算法分类性能的重要因素这一结论。

参考文献

1. 闭小梅, 闭瑞华. KNN 算法综述[J]. 科技创新导报, 2009(14):31-31.

2. 李秀娟. KNN 分类算法研究[J]. 科技信息, 2009(31):81-81.

3. 李航. 统计学习方法[M]. 清华大学出版社, 2012.

4. 周志华, 杨强. 机器学习及其应用[M]. 清华大学出版社, 2013.