K 近邻算法实验报告

王昕毅

（重庆大学软件学院，重庆，401331）

摘要

K 近邻分类算法最初由 Cover 和 Hart 于 1968 年提出，是一个简单而经典的机器学习算法。作为一种非参数的分类算法， K 近邻(KNN)算法是非常有效和容易实现的。 它已经广泛应用于分类、回归和模式识别等。本文对 KNN 算法做了一系列实验，简要介绍了 KNN 算法的思想，原理，实现以及具体的代码，并对实验结果进行了分析。

关键字：K 近邻; 机器学习; 非参数

1. 简介

KNN 分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是: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)的样本中的大多数属于某一个类别，则该样本也属于这个类别。KNN 算法中，所选择的邻居都是已经正确分类的对象。KNN 算法不仅可以用于分类，还可以用于回归。通过找出一个样本的 k 个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。KNN 的分类性能较好。随着理论上的深入研究和分析，现在已在很多领域得到广泛的应用，例如，模式识别、文本分类、图像及空间分类等。本文通过对鸢尾花数据的分类简要展示 KNN 算法的原理，实现步骤以及具体代码。

2. 算法介绍

2.1 KNN 算法三要素

一、距离度量：

(一)曼哈顿距离

$$L_1(\vec{x}_i,\vec{x}_j) = (\sum_{l=1}^{n}|x_i^{(l)} - x_j^{(l)}|)$$

曼哈顿距离又叫街区距离，它就是两个点在各个坐标轴上的分量距离之和。

（二）欧几里得距离

$$L_2(\vec{x}_i,\vec{x}_j) = (\sum_{l=1}^{n}|x_i^{(l)} - x_j^{(l)}|^2)^{1/2}$$

欧几里得距离即空间中两点所连直线的长度，是两点各维度的值的平方和的根

（三）切比雪夫距离

$$L_\infty(\vec{x}_i,\vec{x}_j) = max_{l=1}^{n}|x_i^{(l)} - x_j^{(l)}|$$

切比雪夫距离可以理解成在一个棋盘中，一个棋子每次只走一步，并且可以往旁边 8 个格子的任一个格子移动，按照这样的移动方式从一个点移动到另一个点的最小步数。

本次实验中我们选择欧几里得距离作为距离度量

二、 K 值的选择

K 值的选择对 knn 算法的影响也很大，K 值的物理意义在于在空间上画一个范围，把范围内的点作为近邻点。过小的 K 值对训练集的拟合程度很好但是对一个新的未知的样本可能并不能很好的预测，此时对应模型过于复杂，过拟合。过大的 K 值则使模型过于简单，无法有太好的预测结果，此时欠拟合。在实验中我们选择 K=5 作为近邻的个数。

三、 分类决策规则

分类决策规则即求出未知样本的近邻点之后的分类准则，一般采取多数表决准则，在我们的实验中也采取这一准则。

2.2KNN 算法（线性检索，不包含 KD 树）计算步骤

1）构建训练样本集和测试样本集。

2）设定 k 值。一般先确定一个初始值，然后根据实验结果反复调整至最优。在

本文的实验中将 K 值设定为 5

3）计算测试样本和每个训练样本的距离,本文的实验中如未提起皆采用欧氏距离

4)　选择 k 个近邻样本，选择距离相对较小的 k 个样本作为测试样本的 k 个近邻。

5）通过多数表决准则对预测样本进行分类

2.3 算法特性以及优缺点

优点：

1.简单，易于理解，易于实现，无需估计参数，无需训练；

2.适合对稀有事件进行分类；

3.特别适合于多分类问题(multi-modal,对象具有多个类别标签)

缺点：

1. 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. 计算量较大

3. 实验部分

在这部分内容中我们将具体演示 knn 的算法实现以及测试结果。

3.1 算法描述

输入:训练数据集 A，测试数据 B，k 值

输出：测试集的输出标签类别，正确率

初始化：将训练与测试数据集录入，对于每个数据集将坐标与标签分开存储到两个列表中，

将坐标取出，转为数字。

步骤一：对于每一个测试样本点，依次从训练集中取出样本计算距离，将与当前点的距离以及该点在训练集中的下标一起存储在距离列表中。

步骤二：对距离列表进行排序，前 K 个点即 K 近邻点，通过该点在训练集中的下标找到其标签

步骤三：选出出现次数最多的标签作为样本点的标签

重复以上步骤直到所有训练样本都被分类。计算出分类正确率　，正确率=分类

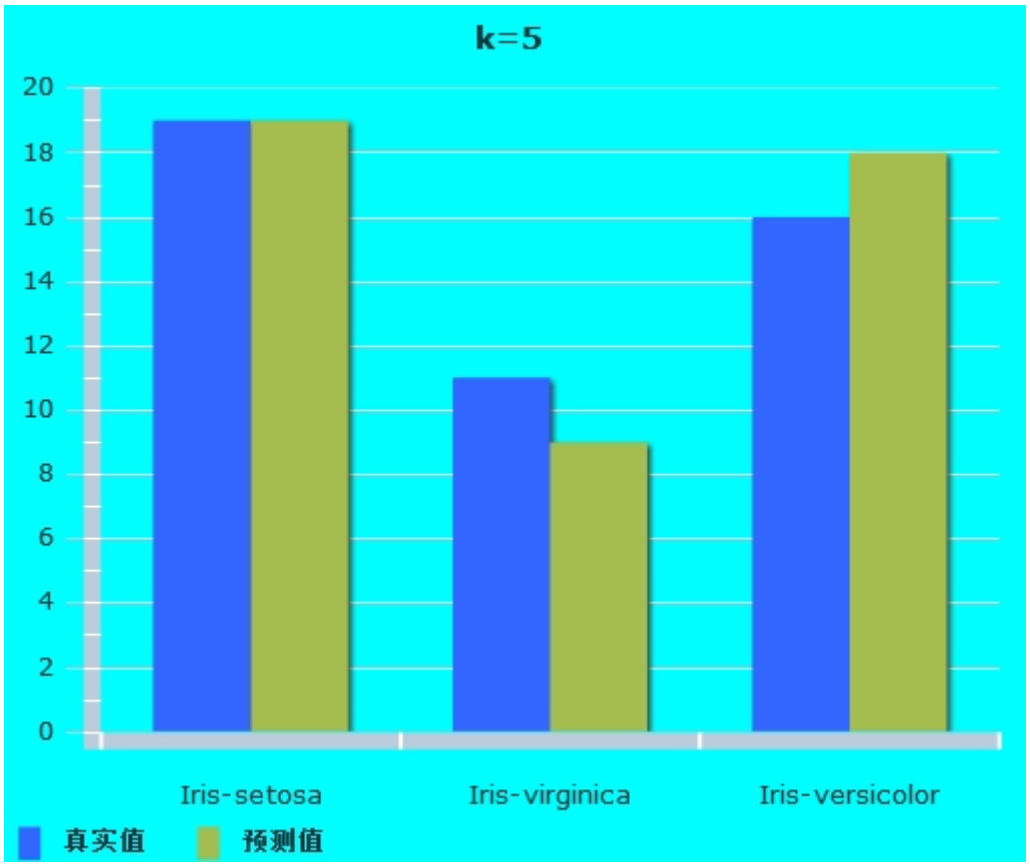
正确的样本数量/分类错误的样本数量

3.2 实验结果

鸢尾花数据集是机器学习中最受欢迎以及被使用最多的数据集之一，它包含了花萼长度，
花萼宽度，花瓣长度，花瓣宽度 4 个属性用于预测鸢尾花属于哪一类。本次实验中我们选取了 104 个数据形成测试集，46 个数据形成了训练集。

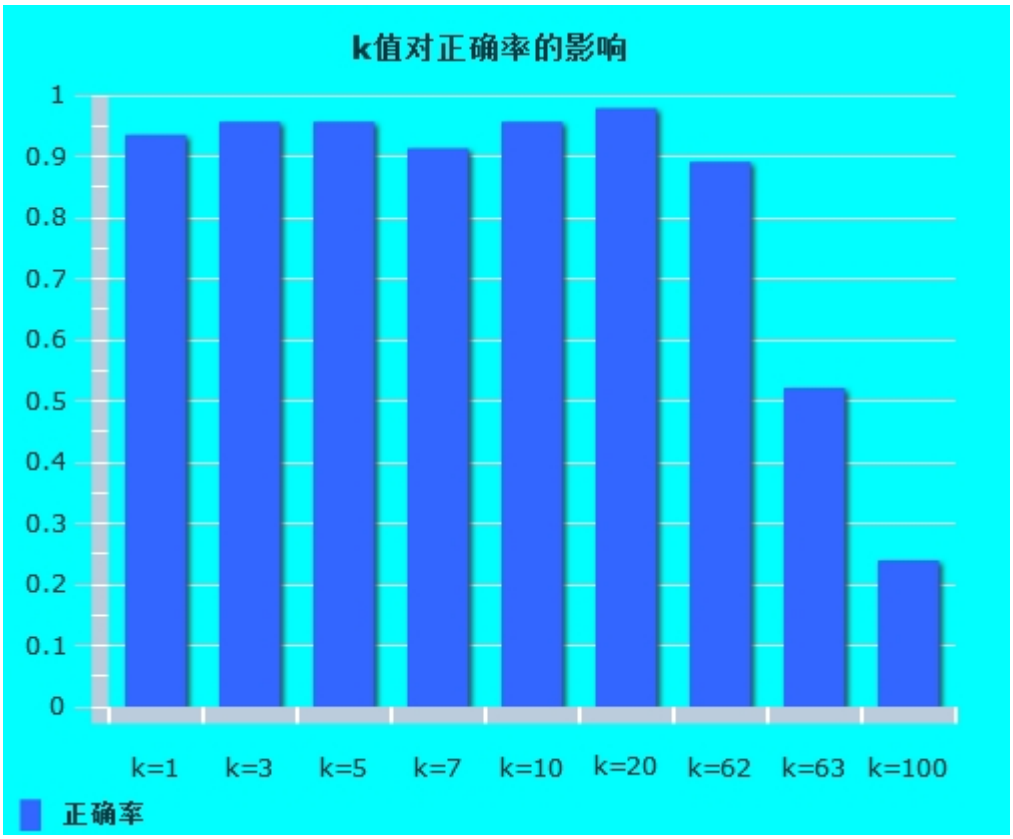
实验结果如下所示

图 1. k=5 时的预测结果



为探究 K 值的选取对实验结果的影响情况，对不同的 K 值进行了测试，结果如下图所示

图 2. K 值的选取对预测正确率的影响



过小的 K 值对应过于复杂的模型，泛化性能不佳，过大的 K 值则对应过于简单的模型，无法很好的拟合数据。由图可见在 K=20 时预测正确率最高，而 K>62 之后随着 K 值的增大分类的正确率逐渐下降。

4. 总结

在这篇实验报告中，我们基本介绍了 KNN 算法的三个要素，采用鸢尾花的数据对 KNN 算法的实验步骤进行了讲解，并对 KNN 算法进行了分析。对不同的 K 值进行实验，得到了不同的预测准确率，得出了选择适当的 K 值是影响 KNN 算法预测准确率的重要因素这一结论。