基于 KNN 的花卉分类技术应用

杨 晨,林国宇
(广州工商学院,广东 佛山 528131)

摘要:KNN 算法在数据分析领域有着重要的应用。文章针对 KNN 算法中参数 k 选择不合理将导致分类结果准确率低的问题,将 K 折交叉验证法应用于 KNN 算法中 k 值的选取,通过 k 值分析图选取最佳 k 值,利用 Python 语言并基于 Sklearn 库实现 KNN 算法。在鸢尾花数据集上的实验表明,该模型是进行花卉分类的有效方法。

关键词:KNN 算法;聚类;花卉分类
中图分类号:TP399 **文献标志码**:A

0 引言

分类算法是人工智能领域研究的重要分支,通过对比样本间的相似度进行分类。分类算法已被广泛应用于智慧园林[1],用于对城市中的各类花卉进行识别和检测,从而有针对性地预防各类花卉病虫害,保护植物的健康生长,有利于加快我国城市的绿色低碳发展。根据植物的生长特性和形态属性来对花朵进行描述是识别植物的高效、简便的方法。

KNN 作为分类算法中最简洁清晰而逻辑明了、快速简单且最高效的分类算法之一[2],在不同领域都得到了广泛应用。本文旨在利用 KNN 算法对植物的花朵进行识别,并在鸢尾花数据集上进行验证。

1 KNN 算法介绍

KNN 算法通常被人们泛称为最近邻居算法或被称为 K 最近邻(K-Nearest Neighbor)分类算法,KNN 算法的一个指导性思想是“近红赤,近墨黑”,根据当前已知的最近邻居来推断出所属类别。KNN 算法有计算简单与便捷、精准度比较高、对异常和数值误差不特别敏感等突出特点。其首次是在 1968 年由 Cover 和 Hart 发表出来,发展到当代在理论上已经相对完善,是一种举足轻重的非参数类的分类算法。非参数类的特性使得该算法仅仅只通过对花卉特征数据进行拟合来分析,而不对特征数据进行参数修正后再进行分类,使得最终的分类结果能更加贴合实际情况。

首先,计算每个待排序训练样本数据与已知训练类别数据之间的最小距离,并尽可能多地找到下一个可能最接近待排序结果的训练样本数据中的前 k 个相邻点。其次,在进行分类或检验后的训练样本数据中包含的训练要素类别数可以仅根据该样本的相邻的各训练要素数据或所属训练要素数据的类别确定[3]。具体实现算法步骤如下:

(1)创建用来训练的样本集合 D。

(2)设定 k 值。确定一个初始值,根据多次重复试验得出最优的结果。

(3)计算得出各个测试样本与每个训练样本间的欧氏距离。从训练样本集合中选出和测试样本的欧式距离最近的样本为作为测试样本的 k 个邻近值。构建一个 n 维空间向量,将样本放置于其对应的 n 维空间 R^n 中,根据欧式距离定义,把任意的样本 x 表示为特征向量 $x=(x_1,x_{2\ldots},x_m)$,则 x_m 为第 m 个特征的值,设任意两个样本 x_i 和 x_j 之间的距离定义为 $d(x_i,x_j)$,其中,$d(x_i,x_j)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$。

(4)选择 k 个近邻样本。将根据公式计算出来的欧式距离从小到大排序,选择欧式距离相对接近的 k 个样本作为测试样本中的 k 个邻近样本。

(5)寻找主导类型。假设 k 个近邻样本为 $x_1,x_2,\ldots,x_k$,其类型标签设为 $c_1,c_2,\ldots,c_t$,将这些类型标签统一归类于标签集 C。根据 k 个近邻的类型采用最大概率法对邻近样本进行分类。该方法中的概率是每一类型的 k 个邻近样本占所有邻近样本的比例,用每种类型在 k 个邻近中的样本数量除以总数 k 来进行计算。将其中具有最大概率的类型作为主导类型。设 $S=\{s_1,s_2,\ldots s_t\}$ 为 k 个近邻中每类别样本数量的集合。设 $\Omega=S_{\mathrm{Max}}$ 为主导类型。

(6)将待测样本归为 Ω 类。重复执行 3,4,5 直到 Ω 趋于稳定。

1.1 KNN 算法的优点

KNN 算法操作简便、易于理解,易于处理多模分类和多标签分类问题(Multi-modal,即研究对象具有多种类型),尤其是对于鸢尾花这类有具体明显特征的植物。同时,当类别体系的变化和训练集的变化时重新训练很便捷(例如:对鸢尾花添加新的训练样本)。不需要应用训练,只需要输入样本集计算机完成分类后,当再输入新的样本时就会识别并自动分类

作者简介:杨晨(2001—),男,广东湛江人,本科生;研究方向:数据分析。

出来。

1.2 KNN 算法的不足

KNN 算法中的参数 k 表示着 k 个距离最近的样本,结果的正确与否与 k 的取值有着很大的关系,如图 1 所示。当 $k=3$ 时,图中方块应该为三角形,而 $k=5$ 时图中的方块应该为圆形,所以参数 k 不同的取值有着不同的结果。选取过小的 k 值,使得训练对象数量太少,导致分类结果受到噪声影响大大增加,同时模型会变得复杂化,容易出现拟合结果过拟合的现象,但 k 值如果过大,则会引入过多的噪声样本,导致分类结果的准确度降低,同时模型又变得单一化,使得结果出现欠拟合的现象。总之,k 值选择太小导致邻居数量太少,会降低分类精度,还会进一步放大噪声数据和干扰信号数据。如果 k 的值太大,那么在样本分类中真正属于该类的数据样本较少,会导致实际上不相似的数据也会包含在内,导致噪声增加,分类效果降低[2]。由于传统上的 KNN 算法往往需通过反复实验和计算来准确选择合适的 k 值,本文主要采用交叉验证法选取最优值的 k 值。

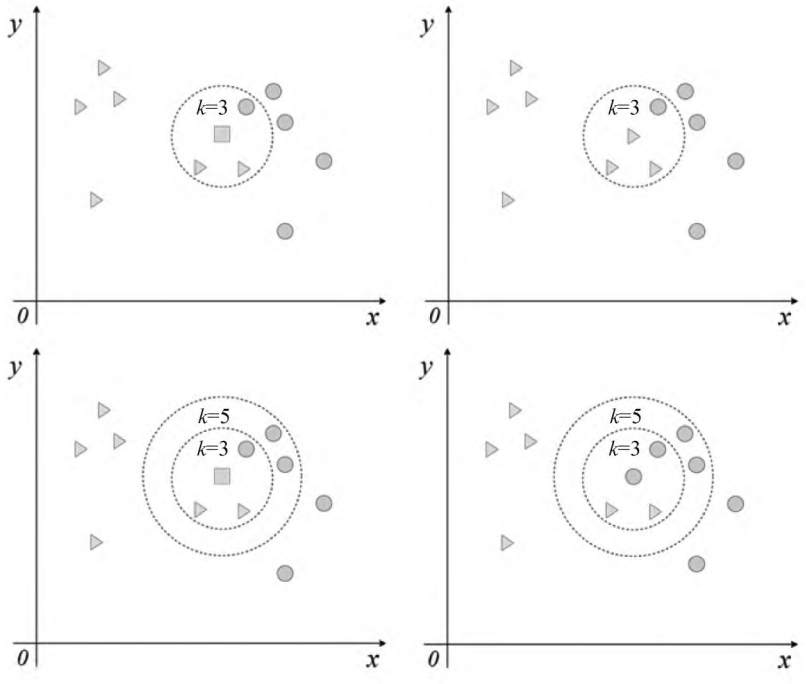

图 1 不同 K 值下 KNN 分类效果

2 利用 KNN 算法实现植物分类

本文通过交叉验证的方法寻找最优 k 值,利用植物花朵的特征来判断类别。使用 Python 语言实现该实验,数据集采用 Sklearn 官方提供的鸢尾花相关数据集。

2.1 K 折交叉验证的优点

K 折交叉验证[4]可以解决植物花朵数据集的数据量不够大的问题和 KNN 算法中 k 值参数寻优的问题,该方法能够有效提高模型评估的准确性,对于样本外数据有更高的准确率,从而最大程度地发挥原始的所有样本数据的作用。

2.2 k 值寻优

KNN 算法中参数 k 的取值对结果有显著的影响,本文考虑通过 K 折交叉验证的一种方法来准确选取 k 的值,该验证方法将原始的所有样本数据分成 n 份相同容量的样本子集(“折”),随机选取其中一份作为测试集,接着拿其他的 $n-1$ 份样本数据组成训练集训练模型,为样本进行训练,接着计算各个模型在测试集上的均方误差 MSE_i ,将 n 次 MSE_i 取算术平均后得到 CV_n 。

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

根据上式计算出错误率。笔者先取较小的 k 值,接着逐渐增加 k 值的大小,再通过计算验证集合的方差,根据方差数值和错误率绘制分析图,错误率最低处对应的 k 值即为最优解,选取对应 k 值为 KNN 算法中的 k 个距离最近的样本。

2.3 算法实现

本文使用经典的鸢尾花数据集,数据集包含了鸢尾花、变色鸢尾花、弗吉尼亚鸢尾花 3 种不同的花型品种中鸢尾花的花瓣长度、花萼长度和花萼宽度还有花瓣宽度这 4 项特征属性值,鸢尾花形态如图 2 所示。

使用 Python 语言中的 Sklearn 库对 KNN 算法进行模型搭建,本文使用 KNN 算法进行花卉分类的算法步骤如下。

图 2　鸢尾花形态

(1)计算错误率:读取 Sklearn 官方提供的鸢尾花相关数据集并循环设置 k 值,其中 $k \in [1,31]$,根据 K 折交叉验证的方法计算出错误率。

(2)绘图:利用错误率与 k 值绘制分析图,如图 3 所示。

(3)选取 k 值:通过分析图可以看出当 k 接近 11 时,错误率达到最低值。对比实验采用不同的 k 值对结果进行分析。

(4)拟合数据:在 KNN 算法中存在一个重要的权重参数 weights。算法模型本身对应是 uniform 与 distance。uniform 参数主要表示距离的大小而与权重参数无关。distance 参数表示出来的数据是权重大小和距离的大小成反比,权重越小,距离预测目标越远。本文分别采用这两种方式来构建 KNN 分类器。

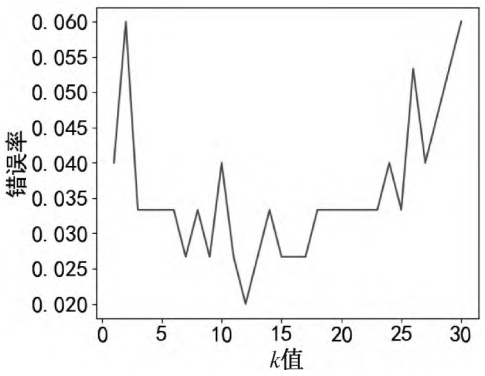

图 3　k 值分析

(5)结果如图 4 所示。图 4 为不同 k 值不同权重下的分类结果。

2.4　结果分析

通过表 1 看出,本文实现的 KNN 算法在 k = 11 时,无论在权重为 distance 的情况下还是 uniform 的情况下,其准确率都明显高于其他的取值。当 k = 11 时,色彩范围边界整体上要比其他取值时更平滑,并在两种权重下都能够取得较好的效果,说明本文算法具有有效性和鲁棒性,如图 4 所示。

表 1　不同 k 值下 KNN 算法算出准确率

权重	k=11	k=3	k=5
distance	74.00	68.00	72.00
uniform	76.00	66.00	68.00

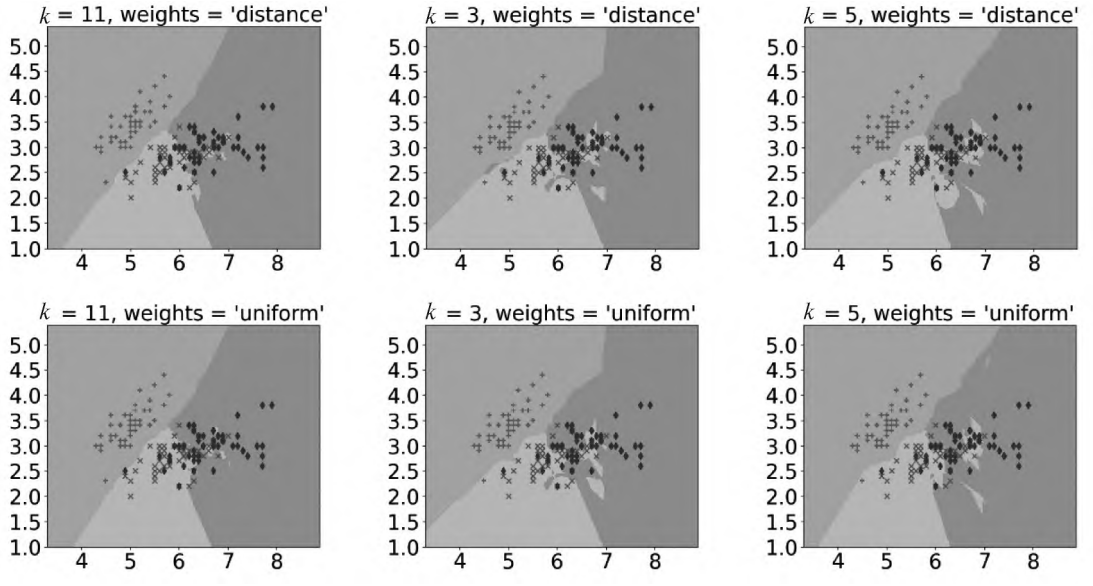

图 4　不同 k 值不同权重下的鸢尾花分类结果

3　结语

在 KNN 算法中,结果的准确性与参数 k 的取值密切相关。为了选取最优的 k 值,本文采用了 K 折交叉验证方法。通过根据鸢尾花的特征描述对花卉进行品类划分,本文在 KNN 算法的不同权重策略下进行实验。实验结果表明,使用 K 折交叉验证方法选择参数 k 时,准确率更高。然而,本文的研究还存在一些不足。尽管准确率有所提高,但最高仅达到 76%,仍有较大的改进空间。另外,K 折交叉验证方法在选

(下转第 73 页)

Application of improved gradient descent method in inertial positioning attitude estimation of underground pipeline

Zhang Nan, Cui Houkun, Xu Weizhou
(State Grid Jiangsu Electric Power Design Consulting Co., Ltd., Nanjing 210008, China)

Abstract: Aiming at the disadvantages of low efficiency and uncertainty of determining the optimal parameters in attitude estimation when traditional algorithms are used for inertial positioning of underground pipeline, a gradient descent method with adaptive parameter adjustment is proposed. This method fuses the data of accelerometer, gyroscope and magnetometer, and automatically adjusts the gradient descent parameters to update the quaternion attitude information. The test results show that the accuracy of roll angle and pitch angle estimated by this method is consistent with the traditional gradient descent method, the accuracy of yaw angle is improved by 11.2%, and the calculation efficiency is improved by 50% compared with the extended Kalman filter method.

Key words: attitude estimation; quaternion; gradient descent method; underground pipeline; inertial positioning technology

(上接第 55 页)

取 k 值时仍需要通过观察来确定,缺乏自适应选择的能力。因此,有效解决上述问题将成为下一步研究的主要方向。未来的研究可以致力于进一步提高分类准确率,并探索自适应选择参数 k 的方法,以改进 KNN 算法在花卉分类中的应用。

参考文献

[1]谢建梅.基于图像处理的农作物病虫害分类算法的研究[J].吉林农业科技学院学报,2021(6):9-13.
[2]窦小凡.KNN 算法综述[J].通讯世界,2018(10):273-274.
[3]SUN J, DU W, SHI N. A survey of KNN algorithm [J]. Information Engineering and Applied Computing, 2018(1):10.
[4]BENGIO Y, GRANDVALET Y. No unbiased estimator of the variance of K-Fold cross-validation[J]. The Journal of Machine Learning Research,2004 (5): 1089-1105.

(编辑 王永超)

Research on flower classification technology based on KNN

Yang Chen, Lin Guoyu
(Guangzhou College of Technology and Business, Foshan 528131, China)

Abstract: The KNN algorithm plays a crucial role in data analysis. This paper addresses the issue of low classification accuracy resulting from the improper selection of parameter k in the KNN algorithm. To tackle this problem, the K-fold cross-validation method is employed for selecting the optimal value of k in the KNN algorithm. The best k value is determined through an analysis of the k value diagram. The KNN algorithm is implemented using the Python language and Sklearn library. Experimental results on the iris dataset demonstrate that this model is an effective approach for flower classification.

Key words: KNN algorithm; clustering; flowers classification