基于 KNN 的图像分类

崔雨豪

摘　要 图像分类是计算机视觉中的基本问题，新的 Fashion Mnist 代替了原有 Mnist 数据集，其在分类难度上较 Mnist 数据集高出许多，更具有参考价值。本文分别利用 KNN 方法和 CNN 方法对 Fashion Mnist 数据进行分类得到以下结论:1.KNN 的 K 值选择在数据集小时表现不稳定，而在数据集大时 K 的取值对准确率影响甚微。2.K=10 时效果较其他取值更优。

关键词 K 近邻；图像分类；深度学习；机器学习

Image Classification Based On KNN

Abstract Image classification is a basic problem in computer vision. The new Fashion Mnist replaces the original Mnist dataset, which is more difficult in classification than the Mnist dataset, and it has more reference value. In this paper, the KNN method and CNN method are used to classify the Fashion Mnist data and obtain the following conclusions: 1. The K value selection of KNN is unstable in the small dataset, However, the value of K has little effect on the accuracy when the dataset is large. 2. K equal to 10 is better than other values。

Key words k-Nearest Neighbor;Image Classification; Deep Learning; Machine Learning

1　引言

图像分类是计算机视觉学科的基本问题，基于Mnist数据集的图像分类算法由于数据集简单不能作为衡量算法好坏的标准。Xiao[1]等人提出了Fashion Mnist数据集，数据集与Mnist相同都是二值化数据集，但内容更加复杂，是关于时尚物品的剪影。针对图像分类的算法有很多。Dudani[2]等人在 1976 年提出了K近邻算法，该算法利用不同距离对图像向量形式进行直接度量，在K个预选值中进行投票法决定出类别，该方法由于未学习特征且计算十分缓慢，故很少使用。Pavlov[3]等人提出了随机森林算法，这种算法具有极高的准确率，随机性的引入，使得模型不容易过拟合，并且有很好的抗噪声能力。深度学习的崛起让图像分类有了新的进展，Krizhevsky[4]等人利用卷积神经网络在Imagenet大赛中轻松击败传统方法，当然，他是Hinton的学生。Simonyan[5]等人提出了VGG网络结构，这也是神经网络开始向更深发展的里程碑。后面还出现Resnet，Inception，Densenet等一系列著名神经网络，并经常用于其他问题的BackBone。

2 本文算法

2.1 KNN

2.1.1 算法概述

KNN 是一种暴力的方法，没有任何训练过程，直接计算距离最近的向量，不考虑特征情况，但是这种方法效果在二值化数据集是有一定效果，本身二值化就可以看作是一种特征而不需要更多的特征提取。这种方法要想效果好必须数据集大，因为没有训练也就不体现找数据共性的思想，也就没有泛化性一说。问题在于这种方法竟然还很慢，所以用这种算法的人少之又少。

2.1.2 距离度量

欧式距离是以一种二范数距离，源自 N 维欧氏空间中两点 x1,x2 间的距离公式：

$$d = \sqrt{\sum_{i=1}^{N}(x_{1i} - x_{2i})^2} \tag{2.1}$$

曼哈顿距离是一种最直接的距离，即两向量相减：

$$d_{ab} = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n| \tag{2.2}$$

3 实验

3.1 KNN

实验环境：Ubuntu16.04+Python3.6.5

3.1.1 基本 KNN 实验

本文实验测试集均采用 Fashion Mnist 中给定数据集并取 100 张进行测试，数据集为 60000 张，变量为 K 与距离算法，即探寻 K 变量与不同距离算法下 KNN 算法的性能：

K	Distance Method	ACC
1	Manhattan	0.84
2	Manhattan	0.84
5	Manhattan	0.84
10	Manhattan	0.86
20	Manhattan	0.84
1	Euclidean	0.84
2	Euclidean	0.84
5	Euclidean	0.84

10	Euclidean	0.86
20	Euclidean	0.84

表 2.1

发现不管在哪种距离下或者 k 值下，60000 张数据集都带来稳定的表现，当 K=10 时在 784 维度上具有更好的效果。

3.1.2 数据集大小与 K 值稳定性度量

数据集大小不同时，K 的变化会带来不同的 ACC 稳定性变化：

K	Distance Method	Dataset Size	ACC	Variance（*10e-5）
2	Euclidean	60K	0.84	8.8
10	Euclidean	60K	0.86	
20	Euclidean	60K	0.84	
2	Euclidean	10K	0.79	6.6
10	Euclidean	10K	0.81	
20	Euclidean	10K	0.8	
2	Euclidean	2K	0.72	42.2
10	Euclidean	2K	0.77	
20	Euclidean	2K	0.75	
2	Euclidean	0.2K	0.69	168.8
10	Euclidean	0.2K	0.65	
20	Euclidean	0.2K	0.59	

表 2.2

实验可以发现，数据集小时，对 K 值的改变极度敏感，从表中也可以体现出 KNN 算法对数据集过分依赖，当数据集减小时，KNN 算法性能急速下降。

参 考 文 献

[1] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms[J]. 2017:7747-7807.

[2] Dudani S A. The Distance-Weighted k-Nearest-Neighbor Rule[J]. IEEE Trans Systems Man & Cybernetics, 1976, 6(4):325-327.

[3] Pavlov Y L. Random forests[J]. Karelian Centre Russian Acad.sci.petrozavodsk, 1997, 45(1):5--32.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[5] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.