Universe Scientific Publishing
Computer System Network and Telecommunications,计算机系统网络和电信(1) 2019,3
ISSN: 2661-3719(Print) ISBN:2661-3727(Online)

K-Means 算法在图像压缩中的应用

毛红霞
四川大学锦城学院 计算机与软件学院 四川 成都 611731

【摘要】在互联网时代,图像信息被广泛应用于各行各业中。图像的数据量非常大,为了有效地传输和存储图像,对图像进行压缩是十分必要的。K-Means 算法是一种最常用的聚类算法。本文将聚类算法的思路应用在图像压缩应用中,并通过 Python 编程使用 K-Means 聚类算法对图像进行压缩处理。从实验结果可以得出 K-Means 聚类方法确实能够对图像进行压缩处理。

【关键词】K-Means;聚类;图像压缩;

聚类分析是数据挖掘中的一个重要研究领域。它将数据对象分组成为若干个类或簇,使得在同一个簇中的对象比较相似,而不同簇中的对象差别很大。图像信息被广泛应用于各行各业中,图像所蕴含信息量越大,占用的空间也越大,因此在传输和保存图片时,要采用合适的方法进行压缩。图像压缩是指在满足一定质量的条件下,是通过消除冗余数据来实现减少表示图像所需的数据量,便于图像的存储和传输。

本文将聚类算法的思路应用在图像压缩应用中,采用 K-Means 聚类算法对图像进行压缩处理。

1 K-Means 聚类算法

聚类是数据挖掘中的一个非常重要的应用领域,所谓聚类,就是根据相似性原则,将具有较高相似度的样本划分为同一类簇,将具有较高相异度的样本划分至不同类簇。K-Means 算法是一种最常用的聚类算法。K-Means 聚类算法又称为 k-均值算法,以距离作为样本间相似性度量的标准,样本间的距离越小,则它们的相似度就越高,那么它们就有可能是在同一个类簇。该算法接收参数 k,然后将样本点划分为 k 个聚类;同一聚类中的样本相似度较高;不同聚类中的样本相似度较低。该算法的思想是以空间中 k 个样本点为中心进行聚类,对最靠近它们的样本点进行归类。通过迭代的方法,逐步更新各聚类中心,直至达到最好的聚类效果。

算法的步骤如下:

(1)从 n 个样本数据集中随机选取 k 个样本作为初始的聚类中心;

(2)分别计算数据集中每个样本到 k 个聚类中心的距离,并将样本划分至距离最小的聚类中心所对应的类中;

(3)针对每个类别,重新计算它的聚类中心;

(4)重复第 2 步和第 3 步直到聚类中心的位置不再发生变化。

整体来讲,K-means 算法的聚类思想比较简单容易实现,并且聚类效果尚可,是一种简单高效应用广泛的聚类方法。

2 K-Means 算法在压缩图像中的应用

2.1 K-Means 算法压缩图像的原理

图像在计算机屏幕上进行显示时会占用计算机的内存空间,一张图片占用内存的计算公式:图片高度 * 图片宽度 * 一个像素占用的内存大小。每个像素包含的字节数直接影响了图像占用内存的大小。存储不同的色彩模式的图像需要不同的内存大小,具体如表 1 所示:

表 1 不同图像类型每像素占用内存小对比表

图像类型	每像素多少字节	可表示的颜色数量
1 比特数据图(Line art)	每像素 1/8 字节	$2^1=2$
8 比特灰度(Grayscale)	每像素 1 字节	2^8
16 比特灰度(Grayscale)	每像素 2 字节	2^{16}
24 比特 RGB	每像素 3 字节,这是图片中最常用的,如 TIF 格式。	2^{24}
32 比特 印刷色彩模式(CMYK)	每像素 4 字节	2^{32}
48 比特 RGB	每像素 6 字节	2^{48}

图像压缩最基本的原理就是将一些相似的颜色用一种颜色来代替,减少描述颜色的种类,则对应的每个像素可以用少的字节来描述,图像所占用的内存就会减少。使用 K-Means 聚类算法进行图像压缩的基本原理:将一张图像中每个像素进行归类,把归为同一类的像素点的值都用这一类的聚类中心的值来代替,通过这

ISSN: 2661-3719(Print) ISBN:2661-3727(Online)

种方法来减少该图像所占用的内存空间。

2.2 实验过程

将图像导入程序时,需要对图像数据进行预处理。根据图片的分辨率,将图片中的数据点平铺,把每个像素点看成一个3维的样本,一张图片的像素值转换成一个n行3列的数据,其中n=height * width。

使用KMeans聚类算法,设定k值,将图片中所有的颜色值做聚类,找出每个3维像素点对应的聚类中心。

```
# 压缩后图片包含的颜色个数,即为聚类的个数
k = 64
kmeans = KMeans(n_clusters=k, random_state=0)
# 训练模型
kmeans.fit(pixel_sample)
# 找到每个3维像素点对应的聚类中心
cluster_assignments = kmeans.predict(pixel_sample)
```

遍历图像中的每一个像素点,并找出每个像素值对应的聚类中心的像素值,并用聚类中心的像素值替换该像素点的值。

```
# 遍历每个像素点,找到聚类中心对应的像素值
pixel_count = 0
for i in range(height):
    for j in range(width):
        # 获取像素点的聚类中心的索引
        cluster_idx = cluster_assignments[pixel_count]
        # 获取聚类中心索引位置上的像素值
        cluster_value = cluster_centers[cluster_idx]
        #替换像素点的值
        compressed_img[i][j] = cluster_value
        pixel_count += 1
```

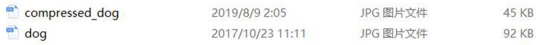
运行代码,图1为压缩前的图片,图2为使用K-Means算法k=64时压缩后的图片。

图1 压缩前的效果

图2 压缩后的效果

2.3 实验结果

由运行结果来看原始的图像是92KB(bird.TIF),而压缩后的图像只有45 K(compressed_bird),实现了对图像压缩的目标。



3 结束语

从实验结果来看,K-Means聚类方法确实能够对图像进行压缩处理,k的取值越小,压缩比例就越大,但压缩后的图像的颜色就越少,就会失去了大量的像素颜色信息,与原图的差异也就越大。K-Means算法算法简单易实现,但需要用户事先指定类簇个数(k的取值),聚类结果对初始类簇中心的选取较为敏感,易陷入局部最优。在实践中,为了得到较好的结果,通常以不同的初始聚类中心多次运行K-Means算法。在所有样本分配完成后,重新计算k个聚类的中心时,对于连续数据,聚类中心应取该类簇的均值。

【参考文献】

[1]胡朝清.K-means算法研究[J].长春工业大学学报(自然科学版),2014(02):25-28.

[2]邹雅莹.基于k-means算法的马田系统研究及其在个人信用评价中的应用[D].南京理工大学,2014.

[4]谭勇,荣秋生.一个基于K-means的聚类算法的实现[J].湖北民族学院学报(自然科学版),2004(01):72-74

[4]徐飞.浅析图像压缩编码方法[J].电脑知识与技术,2010(08):6584-6589

[5]储昭辉.图像压缩编码方法综述[J].电脑知识与技术,2009(18):189-191+194.

[6]图片所占内存.https://blog.csdn.net/heqiang2015/article/details/83618967