

基于 PCA-K-means 的卫星遥感图像的颜色特征提取技术

赵 蓓, 解争龙, 李 红, 李小林
(咸阳师范学院 信息工程学院, 陕西 咸阳 712000)

摘 要: 结合主成分分析(PCA)和 K 均值聚类算法(K-means)的特点, 本文提出了一种对卫星遥感图像进行颜色特征提取的 PCA-K-means 算法. 该算法去除了图像的 R、G、B 之间的相关性, 在动态聚类的基础上, 采用基于区域分类的空间一致性原则合并空间信息, 使得该方法能高效的描述卫星图像的颜色特征. 实验结果表明, 该方法识别性能好, 准确度高, 是对多频谱遥感图像的颜色特征提取的一种有效的方法.

关键词: PCA; K-means; 卫星遥感图像; 颜色特征提取

中图分类号: TP751

文献标识码: A

文章编号: 1000-7180(2012)10-0064-05

Color-Feature Extraction of Remote Sensing Image Based on Principal Components Analysis and K-means

ZHAO Qiang, XIE Zheng-long, LI Hong, LI Xiao-lin

(College of Information Engineering, Xianyang Normal University, Xianyang 712000, China)

Abstract: This article propose a algorithm of PCA-K-means to extraction color feature of remote sensing image. The algorithm realizes the PCA algorithm and K-means algorithm which is suitable for mass data mining. The algorithm move the correlation of R, G and B. Using dynamic clustering method and classify on the basis of region by spatial consistency, the classification algorithm can describe the color feature of the remote sensing image. Experimental results that algorithm of PCA-K-mean has better performance in classification of the remote sensing image.

Key words: Principal Components Analysis; K-means; remote image; color-feature extraction

1 引言

随着高分辨率传感器的应用, 卫星图像的使用越来越广. 卫星图像在公共交通系统、地理信息系统和环境检测系统中发挥着越来越重要的作用. 遥感图像分类是信息提取及处理的关键技术之一. 快速、高精度的遥感图像分类是实现各种实际应用的前提^[1]. 正是因为遥感图像分类方法在卫星遥感的应用中有着非常重要的作用, 很多研究者都在不断试用、研究和改进新的算法, 力图提高卫星遥感图像分类算法的分类效果.

目前对于遥感图形的分类方法主要分为三类:

第一类是基于图像的色彩分类, 如最小距离方法、K-means 方法、最大似然法、神经网络法、主成分分析法; 第二类是基于图像纹理的分类方法, 如共生矩阵法等; 第三类是基于多源数据融合的分类, 多源数据融合就是把多种信息(遥感数据和非遥感数据)按照一定的方式有机的组合在一起的信息模型. 目前遥感图像自动分类方法一般基于图像的统计信息进行分类^[2].

色彩是图像中非常重要的信息. 在图像处理领域中, 色彩信息或称多光谱信息被广泛应用于遥感图像的分类^[3]. 而对于一个分类算法来说, 必须找到一个高效的色彩描述方法. 许多变换都可以将 RGB

色彩空间变换为一个新的色彩空间,而变换都是与实际图像相独立的^[3].因此需要一个更加合理的方法,不仅通过分析目标图像的色彩信息进行编码,而且通过统计分析来对图像编码.

主成分分析(PCA)是一种正交变换,利用二阶的统计信息进行计算^[4].作为一种正交变换,一方面可以减少或消除多波段或多时相之间的相关性对类间的影响,从而提高分类精度^[5].另一方面 PCA 是将多维光谱信息转换为少数几个主成分,这几个主成分包含了大量的图像信息,提高分类的效率^[6].PCA 方法的优势在于数据压缩以及对多维数据进行降维,它将原来众多具有一定相关性的多个指标,重新组合成一组新的互相无关的综合指标来代替原来的指标.

目前,PCA 方法主要应用于人脸识别、性别识别和车牌识别,对图像的特征提取主要在灰度图像方面.对于应用 PCA 进行卫星遥感图像进行颜色特征提取比较少见.而彩色遥感图像一般数据量大,使用 PCA 进行图像特征提取,可大大减少数据量,提高算法的效率.

文中使用 PCA-K-means 算法分析高清晰度敏感图像的色彩信息.该算法将 PCA 算法和 K-means 方法相结合,将其应用于卫星遥感图像.

2 PCA 理论

主成分分析是一种非常有用的数学方法,它是一种最小均方意义上的最优变换,目的是去除输入随机向量之间的相关性,突出原始数据中的隐含特性.它通常被用于人脸识别和图像压缩、特征提取等领域,是在高维数据中寻找模式的一种技术^[7].PCA 是模式识别的一种方法,它以一种强调数据之间的相似和不同的方式表达数据.由于在高维数据中很难发现模式,表达图像就更加困难.

图像特征提取的一个重要的应用领域是在模式识别方面.对于一幅图像,可以通过用线性变换来提取它的颜色特征.这个过程如图 1 所示.

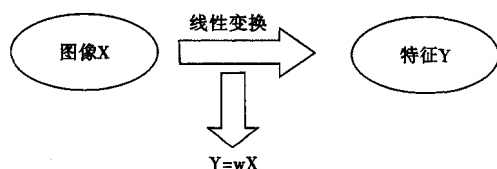


图 1 特征提取原理

这里 w 是一个变换值,可以把它当做基本的变换矩阵.通过此变换来提取图像的特征.

令 x 为表示环境的 m 维随机向量.假设 x 均值为零,即:

$$E[x] = 0$$

令 w 表示为 m 维单位向量, x 在其上投影.这个投影被定义为向量 x 和 w 的内积,表示为:

$$Y = \sum_{k=1}^n w_k \cdot x_k = w^T \cdot x$$

且满足约束条件:

$$\|w\| = (w^T w)^{\frac{1}{2}} = 1$$

主成分分析的目的就是寻找一个权值向量 w 使得表达式 $E[y^2]$ 的值最大化:

$$E[y^2] = E[(w^T \cdot x)^2] = w^T E[x \cdot x^T] w = w^T C_x w$$

根据线性代数的理论,满足式子值最大化的训练应该满足下式:

$$C_x \cdot w_j = \lambda_j w_j, \quad j = 1, 2, \dots, m$$

则使得上述式子最大化的 w ,就是矩阵 C_x 的最大特征值所对应的特征向量.

3 K-means 算法

在图像分类处理的过程中,前期处理完成后,也就准备好了将要参与分类的数据,接下来就要根据这些数据提供的信息进行分类.分类的方法对分类的结有着重要的影响,选择适合的分类方法也是很重要的.分类方式有两种:一种是监督分类法;另一种称为非监督分类法.

常用的监督分类方法有 Euclid Distance 法和 Maximum Likelihood 法非监督分类法以群集为理论基础,通过计算机对图像进行聚集统计分析的方法^[8].根据代分类样本参数的统计特征,建立决策规则来进行分类.不需要事先知道类别特征,把各个样本的空间分布按期相似性分割或合并一群集,常用的非监督分类方法有 K-means 法.

K-means 算法的基本思想是:先选择若干样本为中心进行聚类,对最靠近它们的对象归类.按某种聚类准则使各种样本向各个中心积聚,从而得到初始的分类.通过迭代运算的方法反复修改聚类的结果,逐次更新各聚类中心的值,直至得到最好的聚类结果为止.

计算距离时可采用欧式距离(Euclidian Distance)、曼哈顿距离(Manhattan Distance)和闵科夫斯基距离.

4 基于 PCA-K-means 的遥感图像颜色特征提取算法

本文提出一种将 PCA 算法与 K-means 方法结合的分类方法, 先对卫星图像进行 PCA 变换, 获得 PCA 空间的特征子空间, 然后将训练样本与测试样本分别向该特征子空间投影, 获得识别特征; 利用 K-means 算法进行颜色分类, 最后采用本文提出的空间一致性原则完成卫星遥感图像的色彩分类。

先求 PCA 算法的特征子空间 W_{pca} 。

将图像像素点的 R, G, B 值组成矩阵。设训练样本为 $x_i, i=1, \dots, n, n$ 表示训练样本的数目, 假设颜色分类为 5 类。首先读入样本 $x_i, i=1, \dots, n$, 计算训练样本的均值 avg , 即

$$avg = \frac{1}{n} \sum_{i=1}^n x_i$$

再分别计算各个类的样本均值 $avg_k, k=1, \dots, 5$, 即

$$avg_k = \frac{1}{n_k} \sum_{x_i \in k} x_i$$

其中, n_k 表示属于第 k 类的样本数。

接着, 对样本进行规范化, 即

$$xx_i = x_i - avg, \quad i = 1, \dots, n$$

再计算所有规范化后的训练样本组成的矩阵 X 的协方差矩阵 C , 即

$$C = XX^T$$

然后, 计算协方差矩阵 C 的特征值和特征向量。

将特征值从大到小排序, 将其对应的特征向量组成矩阵, 其转置矩阵就是 PCA 算法的变换矩阵, 也就是特征子空间 W_{pca} 。

用下式把训练样本 X 投影到 W_{pca} 特征子空间中

$$Y = W_{pca}^T X$$

这里, Y 称为基于 PCA 算法的色彩特征。

扫描整个图像, 将训练样本和测试样本的色彩特征进行比较, 采用 K-means 方法进行聚类分析, 距离则采用曼哈顿距离算法。

K-means 算法的步骤是:

- (1) 适当选择 c 个类的初始中心; $c=1, \dots, 5$
- (2) 在第 k 次迭代中, 对任意一个样本, 求其到 c 个中心的距离, 将该样本归到距离最短的中心所在的类,
- (3) 利用均值方法更新该类的中心值;

(4) 对于所有的 c 个聚类中心, 如果利用 (2)、(3) 的迭代法更新后, 值保持不变, 则迭代结束, 否则继续迭代。

最后, 采用空间一致性原则确定该像素属于哪一类, 将该类颜色的中心值赋给它。

由于基于像素的分类方法不包括空间信息, 因此结果会有噪音。而基于区域的分类方法可以克服这个缺点, 得到更好的空间一致性。在该方法中, 基于同一阈域的相邻像素被分到同一个区域中。为避免设计阈域, 文中提出一个简单的、高效的技术来合并空间信息。假设用前面的基于像素分类方法, 已经把像素 (x, y) 分到第 ω 类, 则该像素的临近区域 Ω 中对于每个类的像素数都被计算:

$$\tau_{\omega 1} > \tau_{\omega 2} > \dots > \tau_{\omega i}$$

其中, $\tau_{\omega i}$ 是 Ω 中属于第 ω_i 类的像素数, 把属于每一类的像素数排序, 对前 n 个拥有最多像素的类, 重新分配中心像素到其中距离最近的那一类。

$$\omega_{new} = \arg \min_{1 \leq i \leq n} (dist_{\omega_i})$$

其中, $dist_{\omega_i}$ 是中心像素与第 ω_i 类的中心的欧式距离。

5 实验结果

在实验中, 选取一个 IKONOS 卫星遥感图像如图 2。图像大小为 350×450 像素。训练数据集包括 50 个样本, 选取 5 种颜色, 每类随机选取 10 个训练样本。一个样本像素有 3 个值, 分别对应 R, G 和 B 。在本试验中, 不考虑图像纹理信息。



图 2 原始卫星遥感图像

该图像的 PCA 变换矩阵为:

$$W_{pca} = \begin{bmatrix} 0.7588 & 0.6484 & -0.0618 \\ 0.5654 & -0.6087 & 0.5567 \\ 0.3233 & -0.4573 & -0.8244 \end{bmatrix}$$

一个像素的光谱成分为:

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 0.7588 & 0.6484 & -0.0618 \\ 0.5654 & -0.6087 & 0.5567 \\ 0.3233 & -0.4573 & -0.8244 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

实验中,首先对原始卫星图像进行 PCA 变换,得到图像在 PCA 空间的 R、G、B 图像及其颜色直方图.图 3(a)是原始图像的 R、G、B 图像及其颜色直方图;(b)是经 PCA 变换的图像 R、G、B 图像及其颜色直方图.

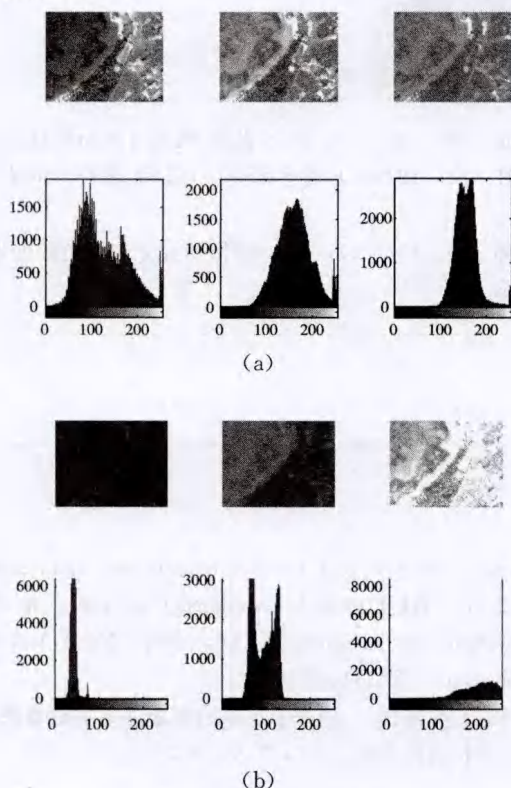


图 3 颜色直方图

其次,在实验中探讨了两种情况.第一,直接在 RGB 空间进行颜色分类,采用最大相似度分类器.第二,进行 PCA 变换后采用 K-means 算法进行颜色分类.为了更加客观的对比变换的效果,两次实验中采用了相同的训练样本.进行 K-means 分类时,

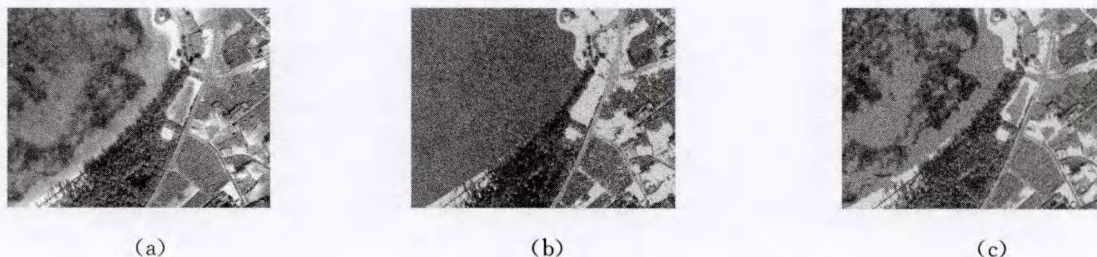


图 4 颜色分类图像

控制迭代次数为 30 次.

实验结果看到,原始图像的 R、G、B 三个直方图有非常相似的形状. G 的分布比 R 和 B 更密一些,这就意味着 RGB 是高度相关的,并且 G 包含更多的信息.而经过 PCA 变换的图像 R、G、B 直方图的分布完全不同,三个成分有非常不同的分布空间,主要分布在 $[30, 70]$, $[80, 110]$, $[120, 200]$,表明通过 PCA 变换去掉了 R、G、B 的相关性.

对基于像素的分类结果,用实验执行了前面所提到的区域分类过程,选定 $n=3$, Ω 是 3×3 的区域.相对于随像素变化的自适应阈域,文中方法的参数更易于调整.

图 4(a)是原始卫星图像;其中(b)是在直接 RGB 的颜色分类图像;采用最大相似度分类器进行颜色分类. (c)是经 PCA-K-means 变换的颜色分类图像.从分类结果来看,采用 PCA-K-means 变换的颜色分类结果比在 RGB 空间效果更好,尤其是对大海、土地和森林的分类效果,更加接近原始图像.这表明,基于主成分空间的颜色分类方法分类准确度高,使用了基于区域的分类的结果比基于像素的分类法好很多.

6 结束语

本文提出了一种基于 PCA-K-means 的卫星图像颜色特征提取算法,将原始卫星图像投影到 PCA-K-means 算法的空间中,提取图像的颜色特征.该算法去掉了原始卫星图像中大量冗余信息,仅保留有用的信息,采用了动态聚类方法.为了合并空间信息,文中提出基于区域分类的空间一致性原则,将基于区域的分类和用光谱主成分获得的图像特征相结合.实验表明,该方法能高效的描述卫星图像的颜色特征,分类准确度高.如何在算法中结合图像的纹理信息来提高分类效果是 PCA-K-means 算法有待继续深入探讨的问题.

参考文献:

- [1] 万朋,王建国,赵志钦,等. SAR 图像点目标的检测[J]. 电波科学学报, 2000,15 (1):55-59.
- [2] 曾生根,王小敏,范瑞彬,等. 基于独立分量的遥感图像分类技术[J]. 遥感学报, 2004,8(2):150-157.
- [3] Dubuisson-Jolly M P, Gupta A. Color and texture fusion; application to imagesegmentation and GIS updating [J]. Image and Vision Computing, 2000(18): 83-832.
- [4] 黄启宏,王帅,刘钊. 改进的基于独立成分分析的图像颜色特征提取算法. [J]. 光电工程, 2007,34(1): 121-125.
- [5] Oja E Principal Components, Minor Components, and linear neural networks [J]. Neural Networks, 1992, 5 (6): 927-935.
- [6] 何国辉,甘俊英. PCA-LDA 算法在性别鉴别中的作用[J]. 计算机工程, 2006,32(19):208-210.
- [7] 余丰宁,蔡启铭. 主成分监督分类及其在水质特征遥感图像识别中的应用. 湖波科学 [J], 1997,9 (3):261-

268.

- [8] 陈华,陈书海. K-means 算法在遥感分类中的应用[J]. 红外与激光工程, 2000,29 (2):26-30.

作者简介:



赵 茜 女, (1971—), 硕士, 副教授. 研究方向为图像处理、软件工程.

解争龙 男, (1961—), 硕士, 教授. 研究方向为信息安全.

李 红 女, (1976—), 博士研究生, 讲师. 研究方向为图像处理.

李小林 男, (1976—), 硕士, 讲师. 研究方向为图像处理、软件理论.

(上接第 63 页)

- [10] Han Jiawei, Pei Jian, Yin Yiwen. Mining frequent patterns without candidate generation[C]// Proceedings of the ACM SIGMOD'00. Dallas, TX, USA: ACM Press, 2000.
- [11] 刘洁,杨路明,毛伊敏,等. 改进的数据流频繁闭项集挖掘算法 [J]. 计算机工程, 2011,37(9): 75-77.
- [12] Show-Jane Yen, Cheng-Wei Wu, Yue-Shi Lee, et al. A fast algorithm for mining frequent closed itemsets over stream sliding window [C]// 2011 IEEE International Conference on Fuzzy Systems. Taipei, Taiwan, 2011: 996-1002.
- [13] Mina Memar, Mahmood Deypir. A block-based ap-

proach for frequent itemset mining over data streams [C]// 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) China: Shanghai, 2011:1647-1651.

- [14] 徐嘉莉,陈佳. 一种基于矩阵的频繁项集更新算法[J]. 计算机应用研究, 2010,27(3): 837-840.

作者简介:

任永功 男, (1972—), 博士, 教授. 研究方向为数据挖掘、图像处理技术等.

李雪兰 女, (1986—), 硕士研究生. 研究方向为数据挖掘.