

红外高光谱数据压缩与评估

Mia Feng

June 5, 2018

云检测: logistic, random forest

仅重述前几次听报告的一些思考。

分类问题解决方法:

- 分类平面: 几何距离 (SVM)、概率分布 (logistic)。
- 分类规则: decision tree (DT), random forest (RF)。

| 派别 | 描述 | 常用估计方法 | 模型类型 |
|-------|-----------|----------------------|-------|
| 频率学派 | 参数是确定的未知值 | 最大化对数似然 | 判别式模型 |
| 贝叶斯学派 | 参数是随机变量 | 最大化后验 (参数的期望作为最优点估计) | 生成式模型 |

Logistic: loss 选择的是交叉熵, 参数估计方法是最大似然。
RF, DT: loss 选择的是互信息 (信息容量) 或者其改进版。
3DVAR、4DVAR: 最大化后验, 可视为生成式模型?

目录

- 红外高光谱数据的特点
- 红外高光谱数据压缩方法
- 基于 KPCA 的红外高光谱数据压缩
- 高光谱数据压缩评估指标
- 问题

数据特点

高空间相关性

每个谱带内某一像素与其相邻像素之间的相似性

高谱间相关性

不同谱带的图像在同一空间位置的像素具有相似性

稀疏性

高光谱数据的高维空间大部分都是空的。

图像自相关函数

$$r_x(l) = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (f(x + l, y) - u_f)}{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2} \quad (1)$$

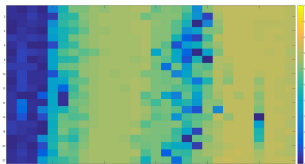
$$r_y(k) = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (f(x, y + k) - u_f)}{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2} \quad (2)$$

自相关系数

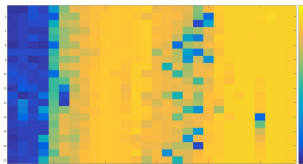
$$\rho_x = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (f(x + 1, y) - u_f)}{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2} \quad (3)$$

$$\rho_y = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (f(x, y + 1) - u_f)}{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2} \quad (4)$$

空间相关性



(a) 行自相关系数



(b) 列自相关系数

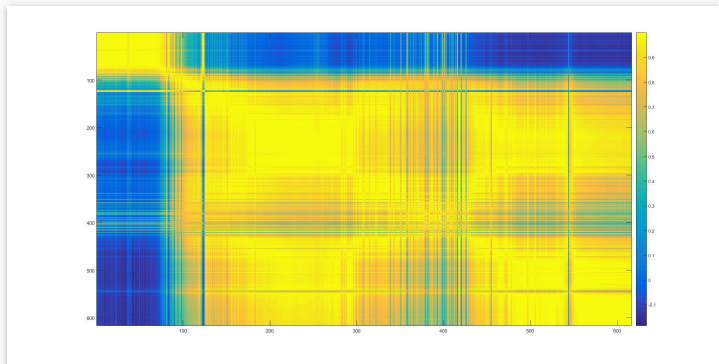
Figure: 自相关系数

$$h(l, k) = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (g(x+l, y+k) - u_g)}{\sqrt{\left(\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2 \right) \left(\sum_{x=1}^M \sum_{y=1}^N (g(x, y) - u_g)^2 \right)}} \quad (5)$$

$$h_{i,j} = \frac{\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f) * (g(x, y) - u_g)}{\sqrt{\left(\sum_{x=1}^M \sum_{y=1}^N (f(x, y) - u_f)^2 \right) \left(\sum_{x=1}^M \sum_{y=1}^N (g(x, y) - u_g)^2 \right)}} \quad (6)$$

$$h_i = \frac{\sum_{x=1}^M \sum_{y=1}^N (f_i(x, y) - u_i) * (f_i(x, y) - u_i)}{\sqrt{\left(\sum_{x=1}^M \sum_{y=1}^N (f_i(x, y) - u_i)^2 \right) \left(\sum_{x=1}^M \sum_{y=1}^N (f_{i+1}(x, y) - u_{i+1})^2 \right)}} \quad (7)$$

谱间相关性



不同谱带间的自相关系数

谱间相关性强于空间相关性。

稀疏性

高光谱数据的高维空间大部分都是空的，所以可以用低维空间去近似表示高维空间，而不会带来较大的误差。推导证明见 page 19-21[7]。

以 IASI 为例，光谱通道计 8461 个，进入同化的通道计 616 个。

The curse of dimensionality

通常是指在涉及到向量的计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象。它描述的是当 (数学) 空间维度增加时，分析和组织高维空间 (通常有成百上千维) 中的数据，因**体积指数增加**而遇到各种问题场景。

降维方法

- 波段选择：特征子空间。
- 特征提取：坐标变换。
- 混合方法：波段选择 + 特征提取。

采用这些降维方法，构建回归方程或者设计码书（**可以看做离散型的映射和连续型的映射？**），可以完成对红外高光谱数据的压缩。

基于预测的压缩技术

一个谱带可以由相邻的谱带预测，其产生的去相关之后的残余误差比较容易压缩。

步骤

- 选取参考谱带：查找准则（等间隔、BH 距离、JM 距离等）；查找算法（最优与次优，SFS、SBS，SFFS、SBFS）
- 建立回归方程：单、双向（是否利用预测的谱带建立回归方程）；多元线性回归。??

基于矢量量化的压缩技术

Idea

编码（码书设计）和解码（码字搜索），分为基于特征选择的矢量量化技术和基于特征变换的矢量量化技术。

以 PCA 为例，编码利用特征向量，解码利用特征向量的逆。也可选择一些子波段用来进行编码（利用高通冗余性）[4]。

基于 KPCA 的红外高光谱数据压缩

与 PCA 的不同：核方法完成非线性映射，可以处理非高斯分布的原始数据。

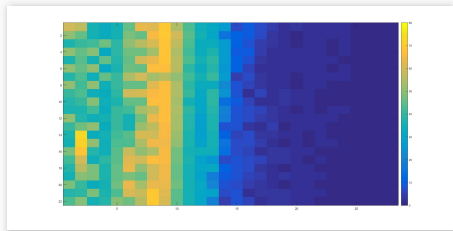
优点：提取非线性映射特征。

缺点：重构的非线性映射算子无法显式得到，需要迭代近似，计算量。且为了迭代近似，需要压缩前的数据。

[步骤]

- 核矩阵计算（非线性映射）：**核函数的选择**。
- 计算特征向量，并按特征值大小排序。
Hint: 为什么 PCA 中 top k 个特征向量重构出的损失最小，证明参见 section 2.12[2]。
- 重构。可以在重构时，对不同 PC 成分指定不同的权重系数。从而分析不同的 PC 可能提取了哪些特征。

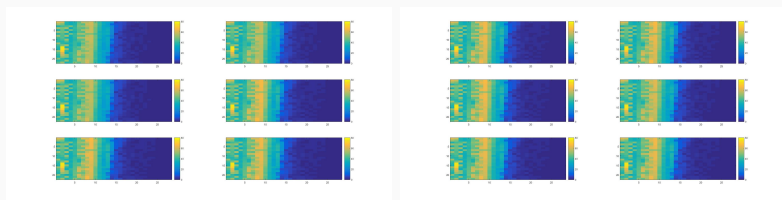
重构结果 [5, 3]



原数据：51（纬度）*120（经度）*616（通道数）

空间上平均后的结果：每个格点代表一个通道在 51*120 上的平均结果
这些图展示的结果没有意义，但我目前只有这些数据。所以这里只说明分析思路。

重构结果——各 PC 在重构中给定权重不同的结果

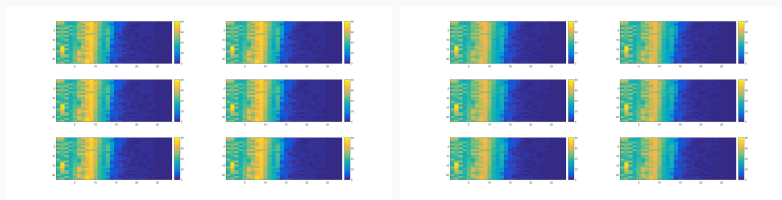


(a) PC 1

(b) PC 2

Figure: PC 重构敏感性

重构结果——各 PC 在重构中给定权重不同的结果

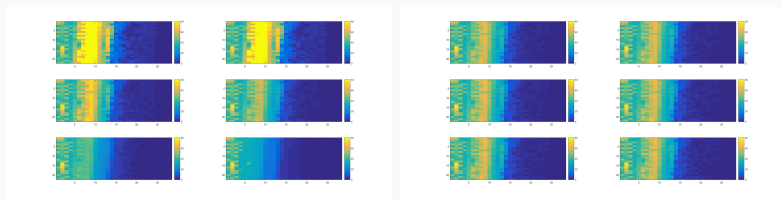


(a) PC 3

(b) PC 4

Figure: PC 重构敏感性

重构结果——各 PC 在重构中给定权重不同的结果



(a) PC 5

(b) PC 6

Figure: PC 重构敏感性

重构结果——各 PC 在重构中给定权重不同的结果

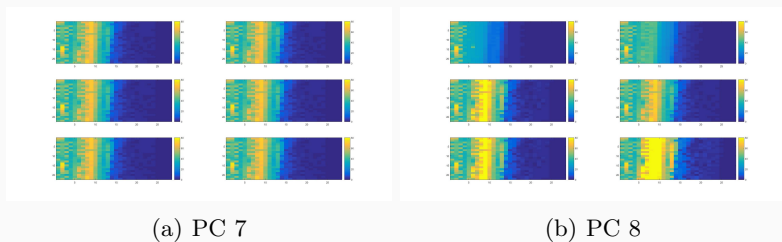


Figure: PC 重构敏感性

重构结果—各 PC 在重构中给定权重不同的结果

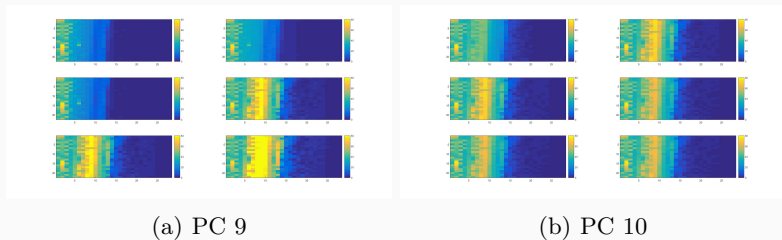


Figure: PC 重构敏感性

可以看到：对不同权重修改后，重构结果差异大的地方所在的谱带不同。得到的 PC 提取到了一些谱带的特征。

高光谱数据压缩评估指标

信息论：信息容量 (Mutual information); 信噪比 (SNR); 峰值信噪比 (PSNR) (page73[7])
压缩比 (CR) (page74[7])
重构正确度：MSE

高光谱数据压缩评估指标

信息论：信息容量 (Mutual information); 信噪比 (SNR); 峰值信噪比 (PSNR) (page73[7])
压缩比 (CR) (page74[7])
重构正确度：MSE

[可参考第六章 [1]]

$$I = H(X) - H(X|Y) \quad (8)$$

$H(X) = - \sum_{x \in X} p(x) \log p(x)$ 为信息熵。特别的，当 $X \sim \mathcal{N}(\mu, \Sigma)$ 时， $H(X) = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \ln 2\pi$ (?? 我推导的与 [6] 公式 2 相比差常数项，请自行验证，如果推导错误请帮我纠错)。

$$I = \frac{1}{2} \log \Sigma_x - \frac{1}{2} \log \Sigma_{X|Y} \quad (9)$$

同化中，

$$I = \frac{1}{2} \log |B| - \frac{1}{2} \log |\Sigma_{X^a|Y}| \quad (10)$$

度量加入观测后，分析场的不确定性被减少了多少。注意，上式第二项的方差不是 3DVAR 公式中的 R。

$\Sigma_{X^a|y}$ 估算 [6]

这里公式的表述是我自己改了一下，请自行对照同化公式再验证一下。不确保我对同化公式的理解一定正确


$$\Sigma_{X^a|y} = B - BK^T (KBK^T + R)^{-1} KB \quad (11)$$

其中 K 为权重函数矩阵，可由 RTTOV 算出。

问题

- 数据集的选取：有云无云？台风与一般天气？海上陆地??
- 权重函数矩阵的计算：RTTOV?
- 分波段分析：数据集标注？

参考文献

-  数学之美.
人民邮电出版社, 2014.
-  Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
The MIT Press, 2016.
-  Sebastian Mika, Alex Smola, and Matthias Scholz.
Kernel pca and de-noising in feature spaces.
In Conference on Advances in Neural Information Processing Systems II, pages 536–542, 1999.
-  V Pellet and F Aires.
Dimension reduction of satellite observations for remote sensing.
part 2: Illustration using hyperspectral microwave observations.
Quarterly Journal of the Royal Meteorological Society,
142(700):2670–2678, 2016.
-  Quan Wang