数据分析中等变非膨胀算子的拓扑理论

刘健

拓扑与几何技术创新中心
河北师范大学

September 22, 2021

目录

认识论

道可道，非常道；名可名，非常名。——老子

我们所了解的世界是我们观测到的世界，而非世界本身。我们的认识不是一成不变的。随着科学技术的发展，认识世界的方法和角度不断丰富，我们的认知更加接近世界的本源。

认识主体和认识工具

- 显微镜的发现让人们观测微观事物，望远镜可以帮助人们观测宏观事物。
- 天气现象综合观测仪是一种智能型多变量传感器，它由一个散射能见度仪，一个降水检测系统传感器以及温度、湿度、风向、风速等传感器组成。
- X射线探测黑洞的存在。当周围物质被它强大的引力所吸引而逐渐向黑洞坠落时，就会发射出强大的X射线，形成天空中的X射线源。
- 皮肤病灶图像，专业医生能看出来，但是普通人没有这个诊断的能力。

观测工具

- 传感器直接测量：对力、电、光、温度、声音等信号的获取——数据。
- 探测工具间接测量：金属探测仪、雷达探测、粒子加速器——通过作用获取数据。

“Data cannot be studied in a direct and absolute way. They are only knowable through acts of transformation made by an agent.”

数学模型

1. 数据表示为拓扑空间上的函数——函数空间。数据在某种意义下具有一定的稳定性，这是由事物特征决定的。
2. 对数据的认识包括数据本身以及对数据的作用——（数据，探测子）。
3. 探测子需要保证数据一定不变性，取函数空间上的等变算子群。
4. 数据的相似性由探测子的作用效果决定。

函数空间

设X为拓扑空间，数据空间Φ指的是所有有界函数$\phi : X \to \mathbb{R}$构成的空间。在Φ中，定义距离

$$D_\Phi(\phi_1, \phi_2) := \|\phi_1 - \phi_2\|_\infty,$$

并假设在该距离诱导的拓扑下，Φ是紧的。

函数空间

Example

一个图片数据可以看成是函数

$$f : X = \mathbb{R} \times \mathbb{R} \to \mathbb{R},$$

其中$f(x)$表示该位置处的像素值。两张图片直接的距离为

$$D(f_1, f_2) = \sup_x |f_1(x) - f_2(x)| / \|x\|$$

其中x跑遍$\mathbb{R} \times \mathbb{R}$中所有像素点。注意到$\|x\|$由$X$中的伪度量

$$D_X(x_1, x_2) = \sup_{f \in \Phi} |f(x_1), f(x_2)|$$

给出。

函数空间

注记

1. 上述例子中拓扑空间X上没有取欧式度量，反而取一种特殊的伪度量。
2. 若Φ是紧的，且X是完备的，则X是紧的。
3. 伪度量在数据分析中比较常见，降维过程就需要构造一个伪度量。并且该降维以后，将一个伪度量空间变成一个度量空间。

等变性

动机：算子的等变性可以算子作用保持数据的对称性。

1. 首先，CNN中卷积算子变换在某种意义下（例如关于平移）本身就是等变的。
2. 对于图片、音频、时间序列数据，平移、旋转、对称、形变都具有不变性。
3. 在CNN中，鉴于上述两点，将这些对称性作为CNN模型的先验知识放进模型中，可以大大提高算法的效率和稳定性。

等变性——群作用

设X是拓扑空间，Φ为如上定义的函数空间，并记$\mathrm{Homeo}_{\Phi}(X)$为所有的同胚映射$g : X \to X$使得

$$\phi \circ g, \phi \circ g^{-1} \in \Phi, \quad \forall \phi \in \Phi.$$

设$\mathrm{Isom}(X)$是关于伪度量D_X的等距同胚组成的空间。则有

Proposition

$$\mathrm{Homeo}_{\Phi}(X) \subseteq \mathrm{Isom}(X).$$

等变性——群作用

设G为$\mathrm{Homeo}_{\Phi}(X)$中等变变换组成的子集，则G是一个拓扑群，其拓扑由下面伪度量给出

$$D_G(g_1, g_2) := \sup_{\phi \in \Phi} D_\phi(\phi \circ g_1, \phi \circ g_2), \quad g_1, g_2 \in G.$$

注意到若G是完备的，则它是紧的。此时(Φ, G)成为感知偶，群G作用可以被学习或者用作为模型的先验知识。

算子GENEOs

GENEOs (group equivariant non-expansive operators)指的是群等变非膨胀算子。

设有感知偶(Φ, G), (Ψ, H)和群同态$T : G \to H$. 映射$F : (\Phi, G) \to (\Psi, H)$称为关于$T$等变的若

$$F(\phi \circ g) = F(\phi) \circ T(g), \quad \phi \in \Phi, g \in G.$$

特别地，若取$(\Phi, G) = (\Psi, H)$和$T = \mathrm{id}$，则F就是我们通常说的G-等变映射。

算子GENEOs

设$F : (\Phi, G) \to (\Psi, H)$是关于$T$的等变算子，若

$$D_\Psi(F(\phi_1), F(\phi_2)) \leq D_\Phi(\phi_1, \phi_2), \quad \phi_1, \phi_2 \in \Phi,$$

则称F是非膨胀的。

非膨胀条件是为了保证算子的有界性和空间的紧性。

算子GENEOs

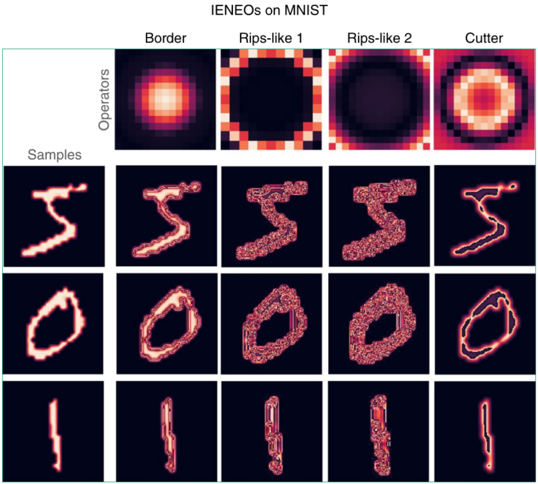

Figure 1: MNIST数据集上作用的等距等变非膨胀算子

算子GENEOs

设$\mathcal{F}^{all} := \mathrm{GENEO}((\Phi, G), (\Psi, H))$为关于同态$T : G \to H$的所有$(\Phi, G)$到$(\Psi, H)$的GENEOs，我们希望$\mathcal{F}^{all}$是紧且凸的。

- 紧性保证GENEOs空间可以被有限逼近。
- 凸性保持GENEOs空间可以由部分算子凸组合张成。

对满足上述条件的GENEOs空间，可以找到一个极小算子空间来生成该空间。

算子GENEOs

Theorem

若Φ和Ψ分别关于D_Φ和D_Ψ是紧的，则$\mathcal{F}^{all}$关于伪度量

$$D_{GENEO}(F_1, F_2) := \sup_{\phi \in \Phi} D_\Psi(F_1(\phi), F_2(\phi))$$

是紧的。

Theorem

若Ψ是凸的，则$\mathcal{F}^{all}$是凸的。

持续同调给出的伪度量

设$\mathcal{F}$是一族GENEOs，则$\mathcal{F}$在函数空间Φ上作用的度量可以很自然地定义为

$$D_{\mathcal{F},\Phi}(\phi_1,\phi_2) := \sup_{F\in\mathcal{F}} \|F(\phi_1) - F(\phi_2)\|_\infty .$$

但是该度量的计算成本是很高的。所以我们寻求一种既合理、又高效的度量。

持续同调提供一种高效、稳定、强不变的伪度量。

持续同调给出的伪度量

伪度量$\hat{d}$是强G稳定的若

$$\hat{d}(\phi_1, \phi_2) = \hat{d}(\phi_1 \circ g, \phi_2) = \hat{d}(\phi_1, \phi_2 \circ g) = \hat{d}(\phi_1 \circ g, \phi_2) \circ g,$$

其中 $\phi_1, \phi_2 \in \Phi, g \in G$。

持续同调给出的伪度量

回顾，对于每个维度k，我们可以得到一个可视化的持续图表，例如

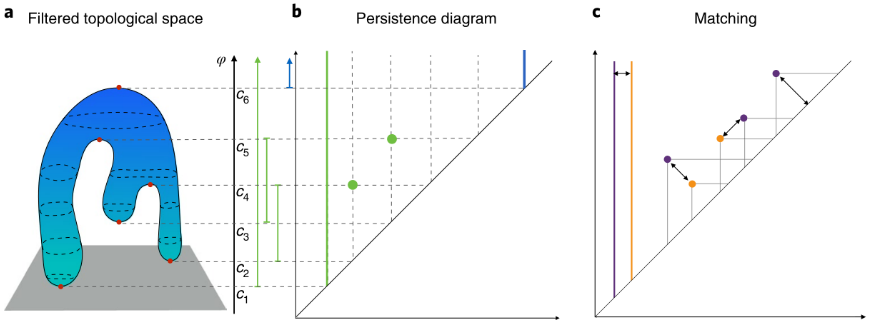

Figure 2: 持续同调

持续同调给出的伪度量

Betti数可以看成一个空间上的实值函数。

Definition 3.12 (Matching distance). Let X, Y be triangulable spaces endowed with continuous functions $\varphi : X \to \mathbb{R}$, $\psi : Y \to \mathbb{R}$. The (extended) *matching distance* d_{match} between β_φ and β_ψ is defined by

$$d_{match}(\beta_\varphi, \beta_\psi) = \inf_{\gamma} \sup_{p \in D_\varphi} \|p - \gamma(p)\|_{\tilde{\infty}}, \tag{3.1}$$

where γ ranges over all multi-bijections between D_φ and D_ψ, and, for every $p = (u, v), q = (u', v')$ in $\bar{\Delta}^*$,

$$\|p - q\|_{\tilde{\infty}} = \min\left\{\max\{|u - u'|, |v - v'|\}, \max\left\{\frac{v - u}{2}, \frac{v' - u'}{2}\right\}\right\},$$

with the convention about points at infinity that $\infty - y = y - \infty = \infty$ when $y \neq \infty$, $\infty - \infty = 0$, $\frac{\infty}{2} = \infty$, $|\infty| = \infty$, $\min\{c, \infty\} = c$ and $\max\{c, \infty\} = \infty$.

持续同调给出的伪度量

故而对于任意k和GENEO集合$\mathcal{F}$，我们可以定义匹配伪度量

$$\mathcal{D}_{\mathrm{match}}^{\mathcal{F},k}(\phi_1,\phi_2) := \sup_{F\in\mathcal{F}} d_{\mathrm{match}}(\beta_k(F(\phi_1)),\beta_k(F(\phi_2))),$$

其中$\beta_k(-)$表示Betti数，$\phi_1,\phi_2 \in \Phi$。

- 持续同调的Betti数计算非常快。
- Betti数之间的匹配距离计算高效，而且

$$d_{\mathrm{match}}(\beta_k(\phi_1),\beta_k(\phi_2)) \leq \|\phi_1-\phi_2\|_\infty.$$

持续同调给出的伪度量

匹配伪度量作为度量$D_{\mathcal{F},\Phi}$的下界是强G稳定伪度量，并且有

$$\mathcal{D}_{\text{match}}^{\mathcal{F},k} \leq d_G \leq D_\Phi.$$

其中

$$d_G(\phi_1,\phi_2) := \inf_{g\in G} D_\Phi(\phi_1,\phi_2\circ g)$$

为自然伪度量。

持续同调给出的伪度量

设$\mathcal{F}$是一族GENOE算子，则可以找到有限子集在匹配伪度量意义下逼近$\mathcal{F}$。

Proposition

设$\mathcal{F}$是一族GENOE算子，对任意$\varepsilon > 0$，存在有限子集$\mathcal{F}^* \subseteq \mathcal{F}$使得

$$|\mathcal{D}_{\mathrm{match}}^{\mathcal{F},k}(\phi_1,\phi_2) - \mathcal{D}_{\mathrm{match}}^{\mathcal{F}^*,k}(\phi_1,\phi_2)| \le \varepsilon, \quad \forall \phi_1,\phi_2 \in \Phi.$$

持续同调给出的伪度量

类似地，对于两个GENOE算子F_1, F_2，我们可以定义算子之间的稳定的伪度量

$$\Delta_{GENEO}(F_1, F_2) := \sup_{\phi\in\Phi} d_{\mathrm{match}}(\beta_k(F_1(\phi)), \beta_k(F_2(\phi))).$$

该伪度量也给出下界$\Delta_{GENEO} \leq D_{GENEO}$，并且具有有限逼近性质。

谢谢!