JSAI 2018
3Pin1-03

潜在変数モデルとナレッジグラフ埋め込みの融合

Incorporating Knowledge Graph Embeddings into Latent Variable Models

東京大学 THE UNIVERSITY OF TOKYO

武石 直也 (理研／東大)　秋元 康佑 (東大)

- 潜在変数モデルなどの**確率モデルのパラメタ推定**に事前知識を活用したい（正則化，制約，etc.）
- 特に**専門家の知識**を有効利用したいが，データに関する知識を直接記述するのは高コスト or 不可能
- 一方，データ生成プロセスの**間接的・あいまいな関係性は比較的容易に記述できる**場合がある
  （例１）データ内の変数同士の正確な関係性は複雑／未知だが，ありうる依存関係の有無は簡単に列挙できる
  （例２）センサの示す物理量の間の依存関係は複雑／未知だが，センサ搭載機器間の関係は簡単に列挙できる

直接的・詳細な関係性

- $\omega = \text{const.}$ if $V > V_{\text{th}}$
- $P \propto V^3$
- $\omega f(\varepsilon) \propto P$
- $\varepsilon = g(P, \omega, T)$　……

→ モデリングにそのまま利用できるが高コスト／未知のため記述しにくい

データ

属性①	V	風　速
属性②	T	気　温
属性③	P	発電量
属性④	ω	回転数
属性⑤	ε	翼　歪
	⋮	

間接的・あいまいな関係性

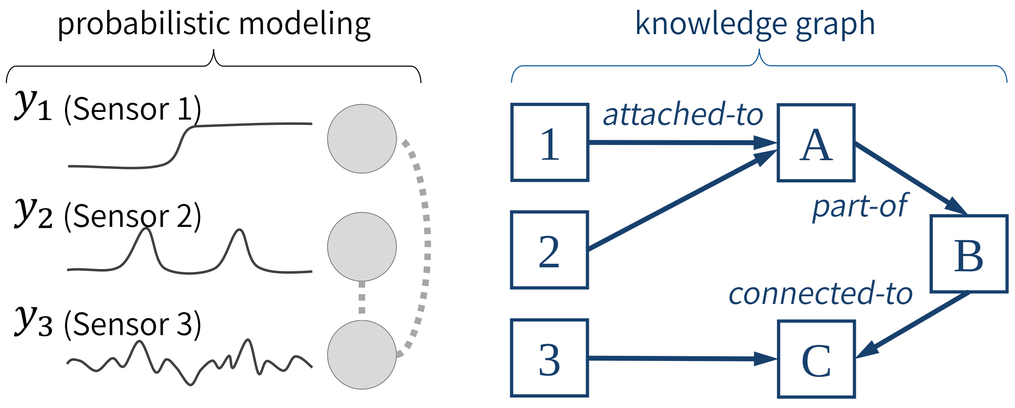
- ① ←(関係あり)→ ③
- ④ ←(関係あり)→ ⑤
- ② ←(関係なし)→ ①
- ⑤ ←(未知)→ ②　……

→ 多くの場合低コストで記述できるが既存手法ではそのまま利用しにくい

Knowledge-Based Distant Regularization

想定する状況

- 潜在変数モデルなどの確率モデルを学習したい
- 利用したい間接的な知識が**ナレッジグラフ**で表せる
- データ内の (少なくとも一部の) オブジェクトや属性に対応するエンティティがナレッジグラフ中にある



提案手法の概要

- ナレッジグラフのエンティティの**連続値埋め込み**[1]**を利用して確率モデルパラメタの事前分布**を用意

関連研究

- グラフ疎性正則化[2]，グラフラプラシアン正則化[3]
- 統計的関係学習の方法による論理ルールの利用[4] 等

例: Distantly Regularized PCA

観測モデル (確率的PCA[5])

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \prod_i \prod_j \mathcal{N}\left(y_{i,j} \mid \boldsymbol{w}_j^\top \boldsymbol{x}_i, \sigma^2\right)$$

- $y_{i,j} \in \mathbb{R}$　オブジェクト i の属性 j の観測値
- $\boldsymbol{x}_i \in \mathbb{R}^d$　オブジェクト i の隠れ変数 (主成分スコア)
- $\boldsymbol{w}_j \in \mathbb{R}^d$　属性 j のパラメタ (因子負荷行列の一部)

提案手法によるパラメタの事前分布

$$p\left(\boldsymbol{w}_j \mid \boldsymbol{e}_j\right) = \mathcal{N}\left(\boldsymbol{w}_j \mid \boldsymbol{f}_\xi(\boldsymbol{e}_j),\, g_\xi(\boldsymbol{e}\_j)\right)$$
$$\text{where} \quad \boldsymbol{f}_\xi \colon \mathbb{R}^k \to \mathbb{R}^d,\ g_\xi \colon \mathbb{R}^k \to \mathbb{R}_+$$

- 属性 $j = 1, \ldots, m'$ に対応するエンティティがナレッジグラフ中にあるとする
- $\boldsymbol{e}_j \in \mathbb{R}^k$　属性 j に対応するエンティティの埋め込み

最終的な目的関数

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2, \xi, \boldsymbol{e}_{1:m'}, \boldsymbol{e}_{\text{KG}}, \boldsymbol{r}_{\text{KG}})$$
$$= (\text{PCA+}\boldsymbol{W}\text{事前分布の周辺尤度}) + \psi(\boldsymbol{e}_{1:m'}, \boldsymbol{e}_{\text{KG}}, \boldsymbol{r}_{\text{KG}})$$

今後，理論的な解析・保証が必要！

- $\psi(\boldsymbol{e}, \boldsymbol{r})$　ナレッジグラフ埋め込みのスコア

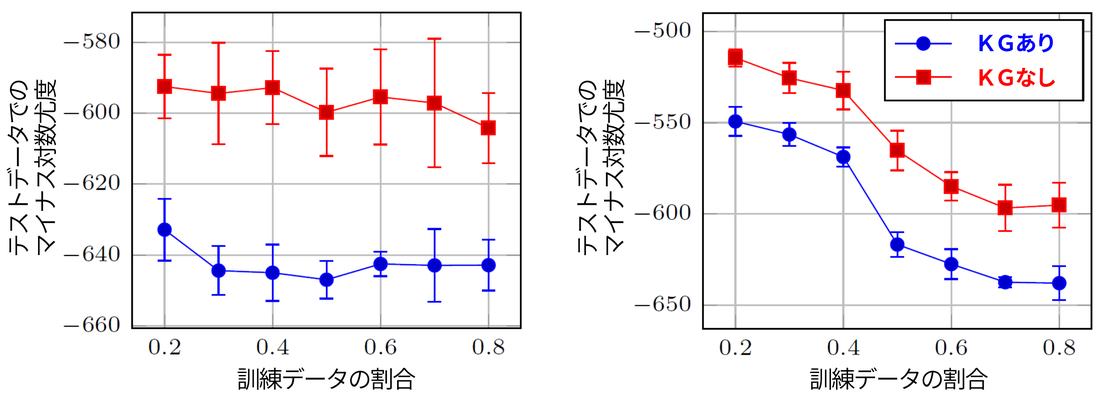
予備実験: Distantly Regularized PCA の汎化性能

データ

- 227の国・地域の月平均降水量 (1901—2015年)[6]

ナレッジグラフ

- 国・地域の地理的知識[7]をナレッジグラフとして利用
  - (Japan, *is-inside*, Asia), (Norway, *is-neighbor*, Sweden) …
- しかし，これは間接的な知識しか与えない
  - 国の地理的関係性は必ずしも気象の関係性を与えない
  - *is-neighbor* は陸の隣接のみ記述 → グラフ正則化は適用し難い



ナレッジグラフによる事前分布を与える場合と与えない場合について，テストデータに関するマイナス対数尤度（小さい値の方がよい）を比較．
ナレッジグラフ埋め込みのスコア ψ としてシンプルな DistMult[8] を使用．
（左）訓練／テスト分割前にデータをランダムに入替　（右）ランダム入替なし

参考文献　[1] サーベイとして Nickel+, *Proc. IEEE*, **104**(1) や Wang+, *TKDE*, **29**(12) 等がある　[2] Huang+, *JMLR*, **12** など　[3] Shahid+, *IEEE J. STSP*, **10**(4) など　[4] Schiegg+, in *AISTATS'12* のほか, Andrzejewski+, in *IJCAI'11* など　[5] Tipping&Bishop, *JRSS B*, **61**(3)　[6] `sdwebx.worldbank.org/climateportal/`　[7] `github.com/mledoze/countries/`　[8] Yang+, in *ICLR'15*

naoya.takeishi@riken.jp