日本オペレーションズ・リサーチ学会
2004年春季研究発表会

階層的GTMに基づく教師付き学習モデル

02103750 筑波大学 後藤正輝 GOTO Masateru
01105930 筑波大学 *香田正人 KODA Masato

1 はじめに

階層的GTM[2]はGTMをノードに持つツリーとして構成される，データ可視化のための教師なし学習モデルである．またGTM(Generative Topographic Mapping,[1])は，高次元のデータの確率分布を，いくつかの潜在変数に置き換えて表現する確率モデルである．

本研究では，階層的GTMが定める潜在変数の確率分布に基づいた教師付き学習モデルを提案する．

2 Generative Topographic Mapping

D次元データ$\boldsymbol{x}$の確率分布$p(\boldsymbol{x})$をL個の潜在変数$\boldsymbol{u}$による表現に置き換えるため，潜在変数空間上の点$\boldsymbol{u}$をデータ空間上の対応する点$\boldsymbol{y}$に写す関数$y(\boldsymbol{u};\mathbb{W})$を定める．このとき$y(\boldsymbol{u};\mathbb{W})$は潜在変数空間をデータ空間に埋め込まれた$L$次元多様体に写す (以下$L=2$とする)．

潜在変数空間上に確率分布$p(\boldsymbol{u})$を定義すると，データ空間上に対応する確率分布$p(\boldsymbol{y}|\mathbb{W})$が生じる．データが多様体上に厳密に分布する状況は非現実的であるので，ノイズを含めた$\boldsymbol{x}$の分布を定義する．

$$p(\boldsymbol{x}|\boldsymbol{u},\mathbb{W},\beta)=\left(\frac{\beta}{2\pi}\right)^{D/2}\exp\left\{-\frac{\beta}{2}\parallel \boldsymbol{y}(\boldsymbol{u};\mathbb{W})-\boldsymbol{x}\parallel^2\right\}$$

データの分布$p(\boldsymbol{x}|\mathbb{W},\beta)$は，潜在変数$\boldsymbol{u}$についての積分として次式で表される．

$$p(\boldsymbol{x}|\mathbb{W},\beta)=\int p(\boldsymbol{x}|\boldsymbol{u},\mathbb{W},\beta)p(\boldsymbol{u})d\boldsymbol{u}$$

$\boldsymbol{u}$についての積分は一般には困難であるため，潜在変数空間上に規則的に格子点$\boldsymbol{u}_i, i=1,\ldots,K$を配置し，$\boldsymbol{u}_i$を中心とするデルタ関数の和として$p(\boldsymbol{u})$を定義する．

$$p(\boldsymbol{u})=\frac{1}{K}\sum_{i=1}^{K}\delta(\boldsymbol{u}-\boldsymbol{u}_i)$$

このとき$p(\boldsymbol{x}|\mathbb{W},\beta)$は次式で得られる．

$$p(\boldsymbol{x}|\mathbb{W},\beta)=\frac{1}{K}\sum_{i=1}^{K}p(\boldsymbol{x}|\boldsymbol{u}_i,\mathbb{W},\beta)$$

データ$\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$が与えられると，EMアルゴリズムによって$\mathbb{W},\beta$の最尤推定値が得られる[1]．

$y(\boldsymbol{u};\mathbb{W})$として，一般化線形モデルを用いる．

$$\boldsymbol{y}(\boldsymbol{u};\mathbb{W})=\mathbb{W}\boldsymbol{\phi}(\boldsymbol{u})$$

$\boldsymbol{\phi}$の要素は,M個の基底関数$\phi_j(\boldsymbol{u}), j=1,\ldots,M$であり，潜在変数空間上に格子状に配置されたガウス分布を用いる．

$\boldsymbol{x}_n$が与えられると潜在変数の事後分布が生成される．

$$R_{in}=p(\boldsymbol{u}_i|\boldsymbol{x}_n,\mathbb{W},\beta)=\frac{p(\boldsymbol{x}_n|\boldsymbol{u}_i,\mathbb{W},\beta)}{\sum_{i'=1}^{K}p(\boldsymbol{x}_n|\boldsymbol{u}_{i'},\mathbb{W},\beta)}$$

全データをまとめて可視化するために，期待値を取ることにより$p(\boldsymbol{u}_i|\boldsymbol{x}_n,\mathbb{W},\beta)$を要約する．

$$E\left[\boldsymbol{u}|\boldsymbol{x}_n,\mathbb{W},\beta\right]=\int p(\boldsymbol{u}|\boldsymbol{x}_n,\mathbb{W},\beta)\boldsymbol{u}d\boldsymbol{u}=\sum_{i=1}^{K}R_{in}\boldsymbol{u}_i$$

3 階層的GTM

階層的GTM $\mathcal{T}$は，各ノードがGTM $\mathcal{M}$で構成されるツリーである．$\mathcal{T}$によって定義される$\boldsymbol{x}$の確率分布は次式で表される．

$$p(\boldsymbol{x}|\mathcal{T})=\sum_{\mathcal{M}\in Leaves(\mathcal{T})}\pi(\mathcal{M})p(\boldsymbol{x}|\mathcal{M})$$

ここで$p(\boldsymbol{x}|\mathcal{M})$は$p(\boldsymbol{x}|\mathbb{W}_{\mathcal{M}},\beta_{\mathcal{M}})$の省略表記である．また$\pi(\mathcal{M})$はモデル$\mathcal{M}$の混合係数であり，次式で再帰的に定められる．

$$\begin{cases}\pi(Root)=1\\ \pi(\mathcal{M})=\prod_{i=2}^{Level(\mathcal{M})}\pi(Path(\mathcal{M})_i|Path(\mathcal{M})_{i-1})\end{cases}$$

ただし$Path(\mathcal{M})_i$はルートノードから$\mathcal{M}$に至るパスを構成するi番目のノードを示す．

$\boldsymbol{x}$が与えられると，潜在変数の事後分布が生成される．

$$p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x},\mathcal{T})=\sum_{\mathcal{M}\in Leaves(\mathcal{T})}p(\mathcal{M}|\boldsymbol{x})p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x},\mathcal{M})$$

4 階層的GTMに基づく教師付き学習

訓練データ$\mathcal{D}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$を学習するため，はじめに$\mathcal{T}$のリーフノードに対応するGTM$\mathcal{M}$の，格子点$\boldsymbol{u}_i^{\mathcal{M}}$についての$\boldsymbol{y}$の期待値を次式で推定する．

$$E[\boldsymbol{y}|\boldsymbol{u}_i^{\mathcal{M}},\mathcal{M},\mathcal{D}]=\frac{\sum_{n=1}^{N}\boldsymbol{y}_n\cdot p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x}_n,\mathcal{M})}{\sum_{n=1}^{N}p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x}_n,\mathcal{M})}$$

データ$\boldsymbol{x}$が入力されると，潜在変数の分布$p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x},\mathcal{T})$が生じる (図1)．このとき対応する出力$\boldsymbol{y}$の期待値を次式で与える．

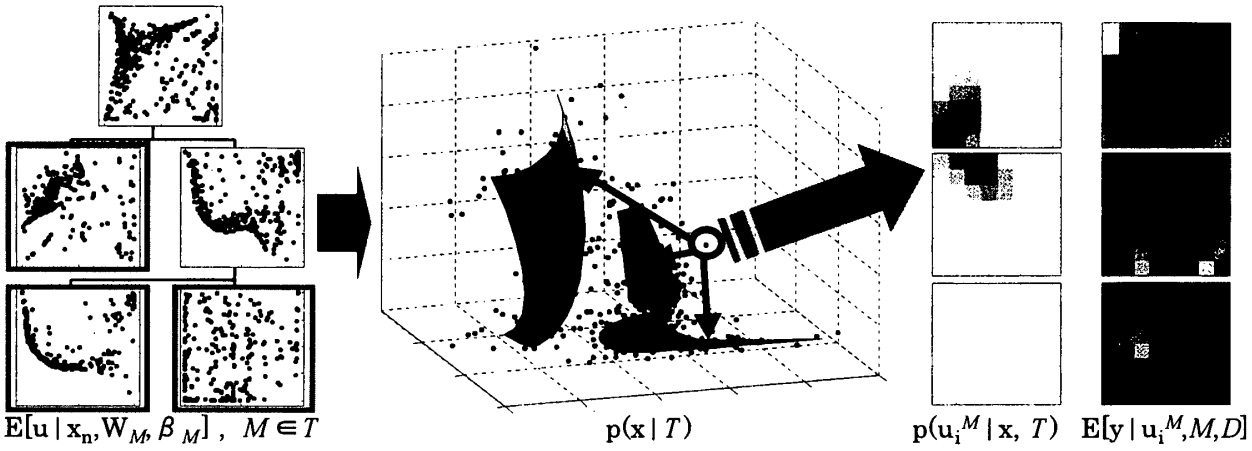

図 1: 階層的 GTM $\mathcal{T}$ は，リーフノードに対応する GTM $\mathcal{M}$ を混合係数 $\pi(\mathcal{M})$ で結合することにより，確率分布 $P(\boldsymbol{x}|\mathcal{T})$ を構成する．データ $\boldsymbol{x}$ が与えられると，潜在変数の事後分布 $p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x},\mathcal{T})$ が生成される．また提案する教師付き学習モデルによって得られた $E[y|\boldsymbol{u}_i^{\mathcal{M}},\mathcal{M},\mathcal{D}]$ を全ての $\boldsymbol{u}_i^{\mathcal{M}}$ についてプロットすることによって，y の期待値のマップが生成される．

$$E[y|\boldsymbol{x},\mathcal{T},\mathcal{D}] = \sum_{\mathcal{M}\in Leaves(\mathcal{T})} E[y|\boldsymbol{x},\mathcal{M},\mathcal{D}]\, p(\mathcal{M}|\boldsymbol{x})$$

ただし $E[y|\boldsymbol{x},\mathcal{M},\mathcal{D}]$ は次式で定められる．

$$E[y|\boldsymbol{x},\mathcal{M},\mathcal{D}] = \sum_{i=1}^{K} E[y|\boldsymbol{u}_i^{\mathcal{M}},\mathcal{M},\mathcal{D}]\, p(\boldsymbol{u}_i^{\mathcal{M}}|\boldsymbol{x},\mathcal{M})$$

階層的 GTM $\mathcal{T}$ のリーフノードに対応する GTM $\mathcal{M}$ それぞれをクラスターとみなし，ファジィ・クラスタリングを規定する．データ $(\boldsymbol{x},y)$ のクラスター $\mathcal{M}$ への所属度を表すメンバーシップ関数 $\mu_{\mathcal{M}}$ を次式で定義する．

$$\mu_{\mathcal{M}}(\boldsymbol{x},y) = P(\mathcal{M}|\boldsymbol{x}),\quad \mathcal{M}\in Leaves(\mathcal{T})$$

また，クラスター $\mathcal{M}$ に属するデータの集合 $C_{\mathcal{M}}$ を次式で定める．

$$C_{\mathcal{M}} = \left\{(\boldsymbol{x},y)\,\middle|\,\mathcal{M} = \arg\max_{\mathcal{N}} \mu_{\mathcal{N}}(\boldsymbol{x})\right\}$$

5 モデル構築の自動化

階層的 GTM $\mathcal{T}$ の構築過程で，GTM $\mathcal{M}$ に属するデータ $C_{\mathcal{M}}$ の分割 $C_{\mathcal{M}_L}, C_{\mathcal{M}_R}$ を定める．

$$C_{\mathcal{M}_L}(i,j) = \{(\boldsymbol{x},y)\in C_{\mathcal{M}}\,|\,d(\boldsymbol{x},\boldsymbol{W}_{\mathcal{M}}\phi_{\mathcal{M}}(\boldsymbol{u}_i^{\mathcal{M}})) \le d(\boldsymbol{x},\boldsymbol{W}_{\mathcal{M}}\phi_{\mathcal{M}}(\boldsymbol{u}_j^{\mathcal{M}}))\}$$

$$C_{\mathcal{M}_R}(i,j) = C_{\mathcal{M}} - C_{\mathcal{M}_L}(i,j)$$

ここで $d(\cdot,\cdot)$ は 2 点間のユークリッド距離を示す．$C_{\mathcal{M}_L}$, $C_{\mathcal{M}_R}$ それぞれに主成分分析を適用し，その結果を近似するように $\boldsymbol{W},\beta$ の EM アルゴリズムの初期値を定めることによって，$\mathcal{M}$ を 2 つの子ノードに分割する．

ノード分割の効果を評価するために，ノードに属するデータの逸脱度 $D(A)$ を定義する．教師信号が連続値の場合，平均二乗誤差を用いる．

$$D(A) = \frac{1}{|A|}\sum_{(x,y)\in A}(y-\mu_A)^2$$

また教師信号がバイナリの場合，エントロピーを用いる．

$$D(A) = -\mu_A\log\mu_A - (1-\mu_A)\log(1-\mu_A)$$

ここで μ_A は A に属するデータの平均であり，$|A|$ は A に属するデータの件数である．分割による $D(A)$ の減少量を分割の効果 ΔD として定め，この値を最大にする i,j に基づいて子モデルの初期化を行う．

$$\Delta D(C_{\mathcal{M}_L}, C_{\mathcal{M}_R}) = D(C_{\mathcal{M}}) - \left\{\frac{|C_{\mathcal{M}_L}|D(C_{\mathcal{M}_L}) + |C_{\mathcal{M}_R}|D(C_{\mathcal{M}_R})}{|C_{\mathcal{M}}|}\right\}$$

ノード分割の停止基準として，$|C_{\mathcal{M}}|$ が $|\mathcal{D}|$ の一定以下になったら分割を停止する minSize と，$|C_{\mathcal{M}}|\cdot D(C_{\mathcal{M}})$ が $|\mathcal{D}|\cdot D(\mathcal{D})$ の一定以下になったら分割を停止する minDev の 2 種類を定める．

6 数値実験

数値実験結果について当日報告する．

参考文献

1. C.M. Bishop, and M .Svensén and C.K.I. Williams. GTM: The Generative Topographic Mapping. *Neural Computation* , Vol. 10, No. 1, pp. 215-234, 1998.
2. P. Tino and I. Nabney. Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, pp. 639-656, 2002.