木構造 SOM による予測クラスタリング

群馬大学理工学府 理工学専攻　　浦澤 毅
群馬大学理工学府 電子情報部門　関 庸一

1　はじめに

本研究では多量の変数を持つ大規模データに対し、木構造を持つクラスタを構築することによる予測手法を提案する。

多くの多変量解析・機械学習の手法では、多変量多事例の全データを用いると計算が膨大となるために、使用する変量やサンプルを事前に選定する作業が必要となる。

そのような場合に対し、自己組織化マップ (SOM, Self-Organizing Maps)[1] を拡張した予測自己組織化マップ (PSOM, Predictive Self-Organizing Maps)[2] は、説明変数の自己組織化を行って生成した類型 (ノード) の分だけを予測に用いるため、全データを用いる予測手法と比べて事前の手間が少なく、容易に多変量の事例に導入できるという利点がある。

しかし、PSOM は説明変数から目的変数への因果関係が複雑である場合は多数の類型が必要となるため、結局収束が遅くなってしまうことが課題であった。

そこで、本研究では木構造 SOM(TSSOM, Tree structured Self-Organizing Maps)[3] の木構造ノードの概念を用い、階層的に用意されたノードにより収束の高速化を図った手法である木構造予測自己組織化マップ (TPSOM, Tree structured Predictive Self-Organizing Maps) を提案する。

2　TPSOM

提案手法を図 1 に示す。PSOM はデータの説明変数 $\boldsymbol{y}_{lj}$ と参照ベクトルの説明変数 $\boldsymbol{u}_{lj}$ 間の距離が最も近いノードについて、目的変数と説明変数の自己組織化を行うことで因果関係を保持したままにノードの学習を行うことを可能とする。提案する TPSOM では自己組織化のための説明変数部分のマッチングを階層毎逐次に行うことで大量のノード探索における負荷の低減を行う。

本手法の評価には実データとして Covertype dataset[4] と人工データを用い、的中率を既存手法と比較してその有用性を報告する。

1: **procedure** TPSOM($\mathbf{X}, T, L, \mathbf{M}, \alpha(t)$)
2:　　**for** $t = 0$ to $T-1$ **do**
3:　　$c^* = 0$;
4:　　$i =$ 一様乱数 on $\{1, \ldots, N\}$
5:　　**for** $l = 1$ to $L-1$ **do**
6:　　$c = \arg\min_{j \in \mathbf{C}_{lc^*}} \|\mathbf{y}_i - \mathbf{u}_{lj}\|$;
7:　　**for** $r \in \mathbf{N}_c^l$ **do**
8:　　$\mathbf{m}_r = \mathbf{m}_r + \alpha(t)(\mathbf{x}_i - \mathbf{m}_r)$;
9:　　**end for**
10:　　$c^* = c$;
11:　　**end for**
12:　　**end for**
13:　　return ($\mathbf{M}$);
14: **end procedure**

- N: サンプル数
- $\mathbf{x}_i$: 第 i サンプルの特徴量 (p 次元)
- $\mathbf{X} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_N^t)^t$: 入力データ ($\mathbf{x}_i = (\mathbf{y}_i^t \mathbf{z}_i^t)^t$)
- $\mathbf{y}_i$: $\mathbf{x}_i$ の説明変数部分
- $\mathbf{z}_i$: $\mathbf{x}_i$ の目的変数部分
- $l = (0, 1, 2, \ldots, L-1)$: 階層番号
- K_l: 階層 l のノード数 ($K = \sum_{l=0}^{L-1} K_l$)
- p: サンプルデータ及び参照ベクトルの次元数
- $\mathbf{m}_{lj} = (m_{lj1}, m_{lj2}, \ldots, m_{ljp})^t$, $(j = 1, 2, \ldots, K_l)$: 階層 l の第 j ノード参照ベクトル (p 次元) ($\mathbf{m}_{lj} = (\mathbf{u}_{lj}^t \mathbf{v}_{lj}^t)^t$)
- $\mathbf{u}_{lj}$: $\mathbf{m}_{lj}$ の説明変数部分
- $\mathbf{v}_{lj}$: $\mathbf{m}_{lj}$ の目的変数部分
- $\mathbf{M} = (\mathbf{m}_{00}^t, \mathbf{m}_{10}^t, \mathbf{m}_{11}^t, \ldots, \mathbf{m}_{1K_1}^t, \mathbf{m}_{20}^t, \ldots, \mathbf{m}_{2K_2}^t, \ldots)$: 参照ベクトルの行列 ($K \times p$ 行列)
- $t = (0, 1, 2, \ldots, T-1)$: 繰り返し数
- $\alpha(t)$, $(0 < \alpha(t) < 1)$: 学習率定数, t の単調減少関数
- $\mathbf{N}_j^l$: 階層 l でのノード j の近傍ノード集合
- $\mathbf{C}_l$: 階層 l のノード集合
- $\mathbf{C}_{lj}$: 階層 $l-1$ のノード j の子ノード集合 ($C_{lj} \subset C_l$)
- $f_t(c)$: $c \in \mathbf{C}_l$ について学習回数 t までに最整合ノードとなったサンプル数

図 1　TPSOM アルゴリズム

参考文献

[1] T. Kohonen: Self-Organizing Maps, Springer (1995)
[2] 南雲, 関, 呉, 矢田, “予測自己組織化マップによる稠密な時系列気象観測データからの短期降雨予測”, 2015 年統計関連学会連合大会, 岡山大学 (2015)
[3] P. Koikkalainen: Tree structured self-organizing maps., In E.Oja and S.Kaski, *Kohonen maps*, Elsvier (1999), pp.121-130.
[4] UCI Machine Learning Repository[URL: http://archive.ics.uci.edu/ml/datasets/Covertype]