最尤共クラスタリングによる二相対応関係の確率モデル化

群馬大学理工学府　海老原 諒介, 関 庸一

1　はじめに

顧客群とブランド群の間の購買関係データなどを理解しようとするとき、二相の対応関係のモデル化が必要となる。観測されるのが顧客の購買の有無のみなら、2 相の関係は単純な二部グラフ構造をもち、共クラスタリング [**?**] の問題となる。

しかし、継続した複数回の購買行動が与えられるような場合は購買頻度等の重み付き二部グラフとなり複雑な関係となる。このような二相間の関係を観測した頻度行列に対して、MDL[**?**] 基準でさらに簡潔なモデル表現を選択し、共クラスタリングするアルゴリズムを提案する。

2　提案手法

顧客 $i(i = 1, \ldots, I)$ とサービス $j(j = 1, \ldots, J)$ が、それぞれ N 個と M 個のセグメント $\boldsymbol{n} \in \{1, \ldots, N\}^I$、$\boldsymbol{m} \in \{1, \ldots, M\}^J$ に整理され、観測データ x_{ij} は各セグメント組合せごとに同一分布に従うものとする。たとえば、顧客セグメント n_i は商品セグメント m_j を確率 $p_{n_i m_j}$ で購入するとすると、ベルヌーイ分布 $x_{ij} \sim \mathcal{B}(p_{n_i m_j})$ となる。生起頻度を $\boldsymbol{X} = \{x_{ij}\}_{j=1,\ldots,J}^{i=1,\ldots,I}$、分布パラメータを $\boldsymbol{P} = \{p_{nm}\}_{m=1,\ldots,M}^{n=1,\ldots,N}$ とする。

ここで、すべてのセグメント間に関係があれば、完全二部グラフ $\mathrm{K_{N,M}}$ で表現されることとなる。しかし、偶然のサービス利用のように閾値 p_{th} 以下の無視できるような弱い関係が含まれており、このような関係は整理して、より単純なグラフで表現できることが期待される場合も考えられる。そこで、閾値以下のパラメータは、所属観測データを統合して求めることで弱い関係を単純化して評価する。このとき、モデルパラメータ数は $l = n \times m - \#\{\hat{p}_{nm} < p_{\mathrm{th}}\} + 1$ となる。

簡潔なモデル表現を選択するために、$f(x|p)$ をデータの確率関数として、(1) 式の MDL 基準を用いる。

$$\mathrm{MDL}(\boldsymbol{P}) = -\sum_{i,j} \log f(x_{ij}|p_{n_i m_j}) + \frac{l}{2} \log IJ \quad (1)$$

第一項は対数尤度であり、モデルの当てはまりの良さを表している。第二項はモデルの複雑さを表している。

アルゴリズムとして図 1 に示す MLBICL を提案する。これは最尤基準に基づき EM アルゴリズムを用いた共クラスタリングである。これを用いて局所解をできるだけ避けるため十分な回数の推定を行い、最良 MDL を与える解を選ぶ。

実データとして、MovieLens データ [**?**] を用いた評価を行った結果を報告する。

procedure MLBICL($\boldsymbol{X}, \boldsymbol{n}, \boldsymbol{m}, K, C$)
  $\hat{\boldsymbol{P}} = \mathrm{Est}(\boldsymbol{X}, \boldsymbol{n}, \boldsymbol{m})$
  $\boldsymbol{P} = \boldsymbol{0}$
  for $k = 1$ to K **do**
    for $i = 1$ to I **do**
      $\hat{n}_i = \arg\max_{n}\ \mathrm{PartLrow}(\boldsymbol{X}, \hat{\boldsymbol{P}}, n_i, \boldsymbol{m})$
    end for
    for $j = 1$ to J **do**
      $\hat{m}_j = \arg\max_{m}\ \mathrm{PartLcol}(\boldsymbol{X}, \hat{\boldsymbol{P}}, \boldsymbol{n}, m_j)$
    end for
    $\hat{\boldsymbol{P}} = \mathrm{Est}(\boldsymbol{X}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$
    $(\hat{\boldsymbol{P}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = \mathrm{Shrink}(\hat{\boldsymbol{P}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$
    $p_{\mathrm{nx}} = \min\{\min_{nm}\{p_{nm} \mid p_{nm} > p_{\mathrm{th}}\}, \max_{nm} p_{nm}\}$
    $p_{\mathrm{bf}} = \max\{\max_{nm}\{p_{nm} \mid p_{nm} < p_{\mathrm{th}}\}, \min_{nm} p_{nm}\}$
    $p_{\mathrm{th}} = \arg\max_{p_{\mathrm{th}} \in \{p_{\mathrm{nx}}, p_{\mathrm{th}}, p_{\mathrm{bf}}\}} \mathrm{MDL}(\boldsymbol{X}, \mathrm{Cut}(\hat{\boldsymbol{P}}, p_{\mathrm{th}}), \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$
    $\hat{\boldsymbol{P}} = \mathrm{Cut}(\hat{\boldsymbol{P}}, p_{\mathrm{th}})$
    if $\max_{nm} |\hat{P}_{nm} - P_{nm}| < C$ **then**
      Break
    end if
    $\boldsymbol{P} = \hat{\boldsymbol{P}}$
  end for
  return ($\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{n}}$, $\hat{\boldsymbol{m}}$, p_{th})
end procedure

- K: 繰り返し数
- $\boldsymbol{n}, \boldsymbol{m}$: 所属セグメント
- C: 収束判定の閾値
- $\mathrm{Est}(\boldsymbol{X}, \boldsymbol{n}, \boldsymbol{m})$: $\boldsymbol{P}$ を推定する関数
- $\mathrm{PartLrow}(\boldsymbol{X}, \boldsymbol{P}, n, \boldsymbol{m}), \mathrm{PartLcol}(\boldsymbol{X}, \boldsymbol{P}, \boldsymbol{n}, m)$: 顧客 n、サービス m の部分尤度を計算する関数.
- $\mathrm{Cut}(\boldsymbol{P}, p)$:閾値 p 以下のセグメント組合せについてプーリングしてパラメータを求め置換する関数.
- $\mathrm{Shrink}(\boldsymbol{P}, \boldsymbol{n}, \boldsymbol{m})$:所属のないセグメントを削除する関数.

図 1: MLBICL アルゴリズム

参考文献

[1] M.Deodhar and J.Ghosh, SCOAL: A Framework for Simultaneous Co-Clustering and Learning from Complex Data, ACM, Trans. on Knowledge Discovery from Data, 4(3), 2010.

[2] Rissanen, J., A Universal Prior for Integers and Estimation by Minimum Description Length, *The Annals of Statistics*,11(2), 416–431, 1983.

[3] F.M.Harper, and J.A.Konstan, The movielens datasets: History and context., ACM Trans. on Interact. Intell. Syst., 5(4), 2016