1I1-2

Deep Boltzmann Machineを用いたデータ融合手法の提案

Data Fusion Method with Deep Boltzmann Machine

新美 潤一郎 *1　　星野 崇宏 *2
Junichiro NIIMI　　Takahiro HOSHINO

*1名古屋大学大学院　　*2慶應義塾大学
Graduate School of Economics, Nagoya University　　Keio University

Nowadays it is difficult for companies to formulate marketing strategy without considering their customers behavior in competing firms. However, even in the big data era, combining large scale data in the company and the general lifestyle survey data is still tough. In this research we created the data fusion model with Deep Boltzmann Machine to combine several datasets collected from different sources.

1. はじめに

企業はマーケティング活動を行うにあたって自社顧客の購買・訪問の頻度や購買額といった情報を活用することに加えて，自社データからは得られない顧客の外部での情報を考慮することも重要である．例として自社の広告接触の有無や競合他社の利用状況といった情報は有益であるものの自社データとして得ることは困難である．これら情報を購買履歴や顧客のウェブ閲覧データに紐付けて分析することは，広告の効果測定や顧客のシェア・オブ・ウォレット (SOW)*1 の把握に有用であると考えられる．

しかしこのような情報は自社の顧客 ID 等の情報を持たずに外部から調査データとして入手することが一般的であり，自社データと紐付けて分析を行うことは困難である．例として自社の購買履歴データ A と外部の調査データ B を組み合わせたデータとして縦軸を観測データ，横軸を得られた変数とした図 1 のような構造が考えられる．共通して得られる J_x 個の変数群 $\boldsymbol{X}$，データ A から得られる顧客の購買頻度等 J_a 個の変数群 $\boldsymbol{Y}_A$，データ B から得られる広告接触等 J_b 個の変数群 $\boldsymbol{Y}_B$ となる．(ここで変数の数 $J = J_a + J_b + J_x$)，関心のある変数群 $\boldsymbol{Y} = \{\boldsymbol{Y}_A, \boldsymbol{Y}_B\}$ はそれぞれデータ A ／ B 一方由来の観測データでしか得ることができないため他方由来の観測データでは欠損値となる．

このようなマルチソースデータにおいて関心のある変数を一消費者について同時に得る手法として本研究で扱う統計的データ融合 (statistical data fusion [Kamakura 97]) がある．データ融合では両データで共通して得られている変数群 X を共変量としてモデリングすることでそれぞれの消費者について Y を同時に得ることを考える．

一方で現在では Deep Learning をはじめとした統計的機械学習の普及が企業にも進み，機械学習を活用

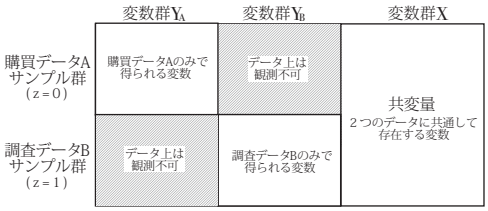

図 1: データ融合で得られるデータ構造の例

した製品・サービスが活発にリリースされている．そこで本研究では Deep Learning の一手法である Deep Boltzmann Machine (DBM [Salakhutdinov 09]) を用いてデータ融合による欠損値の予測を行う．

2. 機械学習における欠損データの扱い

まず欠損データは「1 欠損パターンが完全にランダムである Missing Completely At Random (MCAR)」，「2 データに依存して変数が欠損する Missing At Random (MAR)」，「他の変数に依存して変数が欠損する Not Missing At Random (NMAR)」に大別される．[Rubin 76] 機械学習を含む一般的な統計モデリングにおいて，扱うデータセットが MCAR であれば Case Deletion 等を用いて観測データを削除してモデリングすることができるが，欠損パターンが MAR であるデータセットに対してこのような対処を行うことはバイアスが発生し予測精度の低下につながる．しかしながら一般的な Neural Network モデルでは完全データを用いることが前提となることから MAR の場合を考慮したモデリングは困難であり，実際に多くの機械学習手法においてこのような場合には EM アルゴリズム [Dempster 77] や多重代入法を用いて欠損値を予測した上で投入することが一般的である *2．例外として

連絡先: 慶應義塾大学経済学部 星野崇宏
(e-mail: hoshino@econ.keio.ac.jp)

*1　同一商品カテゴリ内における自社購買の割合

*2　Random Forest [Breiman 01] に代表されるアンサンブル学習においては一部の変数の欠損を考慮したモデルの作成が可能な場合もある

欠損値を含むデータを用いるためのモデルとして，一般的な Feed-Forward Neural Network (FFNN) を拡張した Neural Selective Input Model (NSIM [Lopes 12]) も存在する．NSIM においては FFNN での学習・予測時に各入力変数について欠損のインディケータを導入し，欠損が見られた場合には当該ユニットから隠れ層の各ユニットへの出力を 0 にすることで柔軟にモデルの接続を組み替えながら学習を行う．これによって欠損値を含む MAR なデータセットでもモデリングが可能となる．

しかしながら NSIM ではあくまで欠損データを用いた効率的な予測を目的として一般的な FFNN のみを扱っており，本研究で取り扱うような DBM は用いられていない．DBM は一般的な Neural Network とは異なり特別な出力は持っておらず，モデルに観測データを入力して計算したのちに，全観測変数の予測値を出力する深層学習モデルである．[岡谷 15] DBM ではその構造上，欠損変数の考慮ができれば欠損変数自体の予測が可能となることが考えられる．そこで本研究では DBM で欠損データを扱えるよう拡張し，データ融合のシミュレーションと企業から提供された実データでの試行を実施することでその精度を確認する．

3. 欠損値を持つ DBM モデル

本研究ではデータ融合の実施のため観測データに欠損値を含む DBM モデルを用いる．図 2 に表されるような欠損を持たない一般的な DBM では，解析に用いる J 個の変数を用いた可視ユニット $\boldsymbol{v}$ と隠れユニット $\boldsymbol{h}$ を用いた場合に，

$$p(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h},\theta)p(\boldsymbol{h}) \tag{1}$$

$$= \sum_{\boldsymbol{h}} \left[\prod_{j=1}^{J} p(v_j|\boldsymbol{h},\boldsymbol{\theta})\right] p(\boldsymbol{h}|\boldsymbol{\theta}) \tag{2}$$

として隠れ変数を周辺化することで，尤度関数を

$$L^c(\boldsymbol{\theta}|v_1,...v_N) = \prod_{i=1}^{N} \sum_{h_i} \left[\prod_{j=1}^{J} p(v_{ij}|h_i,\boldsymbol{\theta})\right] \tag{3}$$

と表すことができる．

ここで [星野 09] に従って各観測データに対応する欠損のインディケータ $\boldsymbol{Z} = \{z_1, z_2, ..., z_N\}$ を導入し，ある観測データがデータセット A 由来の場合に 1 を，そうでない場合に 0 を割り当てることにする．これによって欠損を含む観測データの尤度関数を

$$L(\boldsymbol{\theta}) = \prod_{i:z=1}^{N} p(\boldsymbol{Y_A}, \boldsymbol{X}) \prod_{i:z=0}^{N} p(\boldsymbol{Y_B}, \boldsymbol{X}) \tag{4}$$

と表すことができる．よってモデルの尤度関数は

$$L^o(\boldsymbol{\theta}|v_1,...v_N) = \prod_{i:z_i=1} \sum_{h_i} \left[\prod_{j=1}^{J_a} p(v_{ij}|h_i,\boldsymbol{\theta})\right] \left[\prod_{j=1}^{J_a+J_b+1} p(v_{ij}|h_i,\boldsymbol{\theta})\right] p(h_i|\boldsymbol{\theta})$$
$$\times \prod_{i:z_i=0} \sum_{h_i} \left[\prod_{j=J_a+1}^{J_a+J_b} p(v_{ij}|h_i,\boldsymbol{\theta})\right] \left[\prod_{j=1}^{J_a+J_b+1} p(v_{ij}|h_i,\boldsymbol{\theta})\right] p(h_i|\boldsymbol{\theta}) \tag{5}$$

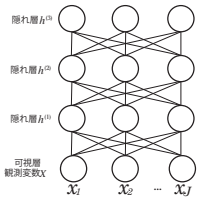

図 2: 通常の DBM

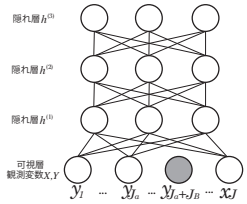

図 3: 欠損値を含む DBM

と表すことができる．これは，欠損値を含む DBM の学習時に，例として変数群 $\boldsymbol{Y}_B$ が欠損している（$z_i = 1$）場合にモデル構造を図 3 のように変更することを表している．

4. データ融合の解析例

本研究では実データを用いる前に擬似データを作成して解析のシミュレーションを実施している．このシミュレーションでは，[Kamakura 00] で提案されている因子分析モデルと同様にモデルの背後に潜在変数として標準正規分布に従う因子 h を仮定し，変数群 $\{\boldsymbol{Y}_A, \boldsymbol{Y}_B, \boldsymbol{X}\} = \{v_1, ...v_J\}$ が全て h に従って生成されているものとする．「一様分布で発生させた係数を用いた潜在変数の線形結合／ 2 次結合」と「発生させた変数が離散値／連続値」の 2 点を組み合わせた 4 つのデータセットを作成している．確率的に $\boldsymbol{Y}_A$ もしくは $\boldsymbol{Y}_B$ を欠損させることでマルチソースデータを作成，DBM でのデータ融合の精度を確認した．

実データへの適用においては，株式会社ビデオリサーチ・インタラクティブより提供された約 1 万 2 千人の PC からの Web 閲覧を記録した Clickstream Data を用いる．国内大手 EC サイト 2 社を対象に自社と競合他社を設定し，データセットに収録されたデモグラフィック情報に加えて購買や閲覧回数をユーザごとに計測した完全データを作成，そのデータに対して確率的に自社または競合他社の情報を欠損させることで，擬似的に購買データと調査データを作成する．このデータセットに DBM を適用することで欠損している変数を予測し，実際の値と比較することで精度を確認する．

各解析結果については当日報告する．

参考文献

[Breiman 01] Breiman, L.: Random forests, *Machine learning*, Vol. 45, No. 1, pp. 5–32 (2001)

[Dempster 77] Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38 (1977)

[Kamakura 97] Kamakura, W. A. and Wedel, M.: Statistical data fusion for cross-tabulation, *Journal of Marketing Research*, pp. 485–498 (1997)

[Kamakura 00] Kamakura, W. A. and Wedel, M.: Factor analysis and missing data, *Journal of Marketing Research*, Vol. 37, No. 4, pp. 490–498 (2000)

[Lopes 12] Lopes, N. and Ribeiro, B.: Handling missing values via a neural selective input model, *Neural Network World*, Vol. 22, No. 4, p. 357 (2012)

[Rubin 76] Rubin, D. B.: Inference and missing data, *Biometrika*, pp. 581–592 (1976)

[Salakhutdinov 09] Salakhutdinov, R. and Hinton, G. E.: Deep Boltzmann Machines., in *AISTATS*, Vol. 1, p. 3 (2009)

[岡谷 15] 岡谷 貴之：深層学習, 講談社 (2015)

[星野 09] 星野 崇宏：調査観察データの統計科学: 因果推論・選択バイアス・データ融合 (2009)