2E3-2

ヘテロなデータに対する統計的学習を用いた傾向スコア推定

内橋堅志　　兼村厚範
Kenshi Uchihashi　Atsunori Kanemura

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

The progress of the ICT technology has produced data-sources that continuously generate datasets with different features and possibly with partial missing values. Such heterogeneity can be mended by integrating several processing blocks, but a unified method to extract conclusions from such heterogeneous datasets would bring consistent results with lower complexity. This paper proposes a flexible propensity score estimation method based on statistical learning with classification, and compared its performance against classical generalized linear methods.

1. 序論

近年、ICT 技術の発展により継続的に生み出されているデータには、情報源が複数であること (ヘテロ性) や計測に必要なコストが異なることで、特徴量の性質が異なり部分的な欠測が生じることが多くある。このようなデータに対して何らかの解析を行う際、欠測を含まない完全データだけを抜き出して行われることが多い。しかし、データが欠測している群と欠測していない群との間に何らかの系統的な差異が存在していた場合、完全データのみから行った解析はバイアスを持つ。加えて、欠測データを除くことによるデータ数の減少なども起こるため、何らかの方法でバイアスがかからないように欠測を埋めて解析を行う必要がある。

ヘテロ性や欠測への典型的な対処法として、変数の性質に応じたモデルを利用することや、平均値補間などの前処理を行う方法がある。これらの手法は欠測の生じ方によって整理されており、ある程度の方法論が確立されている [6, 2] が、本稿では、そのなかでも、傾向スコアによるマッチングに着目し、統計的学習による傾向スコア推定法について描写する。モデリング時には、変数にヘテロ性を仮定し、連続変数だけでなく離散変数が存在するものとする。

本稿では、統計的学習により傾向スコアを推定するモデルとして、従来からよく使われているロジスティック回帰のような線形モデルの拡張として非線形な表現能力を持つモデルを採用し、その効果を検証する。非線形なモデルとして、Gradient Boosting Decision Tree (GBDT) 及び Convolutional Neural Network (CNN) の二つを用いた。これら 2 つにロジスティック回帰を含めた 3 つの学習器それぞれについて、傾向スコア解析においてよく知られた 3 つのデータセットについて傾向スコアを推定し、マッチング精度の比較実験を行う。この際、マッチング精度を比較するために、t 値による評価と Kullback-Leibler 情報量による評価を行う。これによって、非線形な手法が傾向スコア推定を用いたマッチングにおいて有効であることを示し、CNN の有効性について考察する。

連絡先：〒 305–8568 茨城県つくば市梅園 1-1-1 中央第二
産業技術総合研究所　兼村厚範 <atsu-kan@aist.go.jp>

2. 欠測の下での統計解析

欠測値の生じるメカニズムにバイアスが存在する条件下で、バランシングスコアの一つである傾向スコアの値に基いてバイアスの調整を行い、効果量推定のバイアスを低減させる方法論を描写する。

2.1 欠測メカニズム

欠測値は、その欠測メカニズムによって MCAR (missing completely at random)、 MAR (missing at random)、 MNAR (missing not at random) の 3 つに分類される [6]。本稿では MAR の場合を想定するが、本段落では他の 2 つのメカニズムについて簡単に説明する。MCAR は、欠測が単独で統計的に独立に（ランダムに）生じている場合を指す。共変量を $\boldsymbol{x}$、欠測がある効果量を y、欠測が生じたかどうかを示す欠測指標を $r \in \{0, 1\}$ とおくと、MCAR では r が $\boldsymbol{x}, y$ と独立である。なお、効果量 y を *target*、欠測指標 r を *treat* とも呼ぶ。この場合、欠測値を含むサンプルを単純に取り除いて得られる完全データを対象に統計解析を行っても、得られる結論にバイアスは生じないが、データ数が少なくなる問題がある。NMAR は、$\boldsymbol{x}$ のみからは r を推定できず、$\boldsymbol{x}$ を統制した条件下でも r が y と独立にならない、一般の場合を指す。本稿では NMAR は扱わない。

MAR は、観測された $\boldsymbol{x}$ のみで r を説明することができる、すなわち r と y との間に $\boldsymbol{x}$ を条件とした条件付き独立性が成り立つ場合であり、

$$p(r|\boldsymbol{x}, y) = p(r|\boldsymbol{x}) \tag{1}$$

が成立する。この条件下では、欠測の有無を観測されている情報のみ（すなわち $\boldsymbol{x}$）から完全に説明することが可能

である。これは、$\boldsymbol{x}$ を統制した条件下では、欠測の有無 r は欠測しうる効果量 y と独立になるということを示している。このような性質から、後に述べる傾向スコアを用いたモデルを構築することでデータ全体を利用しつつバイアスの無い推定が可能である [3]。本稿で扱うデータセットでは、MAR が成り立っていると仮定する。

2.2 傾向スコア

MAR において、欠測値の有無によらない比較を行うために、両群間で共変量によるマッチングを行う。この時、共変量の次元やデータ数が大きいと、適切なマッチングを行うことは難しいが、共変量の持つ欠測に関する情報を 1 次元に縮約したバランシングスコアを用いることで 1 次元の値に基づくマッチングが可能になる [8]。バランシングスコア b とは、共変量 $\boldsymbol{x}$ の関数であって、

$$p(r|\boldsymbol{x}, b(\boldsymbol{x})) = p(r|b(\boldsymbol{x})) \tag{2}$$

が成り立つような変数であり、これを用いることで、共変量 $\boldsymbol{x}$ 全てを参照することなくバランシングスコア b だけをみれば欠測値の有無 r が条件付き独立となる。

バランシングスコアの一つとして知られるものに傾向スコアがある。傾向スコア $e(\boldsymbol{x})$ は、共変量 $\boldsymbol{x}$ が与えられたときに欠測が生じる確率値であり、

$$e(\boldsymbol{x}) = p(r = 1|\boldsymbol{x}) \tag{3}$$

と定義される。傾向スコアはバランシングスコアである [8] ため、共変量が持つ欠測に関する情報を 1 次元に縮約したものであるとみなすことができ、傾向スコアの値で条件付けることで、欠測値はランダムに生じるとみなせる。

傾向スコアの推定モデルには、一般化線形回帰モデルであるロジスティック回帰が標準的に用いられる [3]。ロジスティック回帰は、各変量の寄与の大きさや p 値の解釈を行うような解析には適しているものの、傾向スコアの値の推定に対しては、高度な非線形性を持つ既存の機械学習手法と比べて高い精度を持っているとは言えない [1]。したがって、本稿で扱う、単に傾向スコアの値を精度良く推定することが求められる問題には必ずしも最適な選択肢ではない。

本稿では、傾向スコア推定に対して、近年よく用いられる機械学習手法である GBDT と CNN を適用し、共変量の近づきをロジスティック回帰と比較することによって評価する。

2.3 マッチング

傾向スコアの値が同じであれば、欠測値の有無に関わらず共変量の分布が等しくなることが理論的に保証されている [7]。これより、$r = 0$ と $r = 1$ の両群から傾向スコアの値が近いペアをマッチングにより作成すると、欠測値はペアのうち欠測がない方の値と近いものになるはずである。本稿では、$r = 0$ と $r = 1$ の両群から最も傾向スコアの値が近いペアを貪欲法によって作成することによってマッチングを行った。

3. CNN による傾向スコア推定

CNN は、convolution 層と pooling 層の重ねあわせと重み共有によってスパースな実装となった多層ニューラルネットワークである [4]。入力の幾何的変化に対する不変性をスパースかつ効率的な重みパラメータ表現として獲得できるため、画像認識分野での利用が主であったが、近年は時系列や言語表現を入力とした場合でもよく利用されており、高い精度を示している [10], [5]。入力変数と比べて低次元なフィルターの重みを学習するため、通常の DNN より計算量が少なく、かつ誤差逆伝搬学習における誤差勾配の発散が起こりにくいことから過学習もある程度抑制されている。また、入力に対する非線形性を持つことから、線形のモデルと比べてはるかに高い表現能力がある。

convolution 層はモデルに検出対象の平行移動やサイズ変化に対する不変性を持たせつつ局所特徴を抽出する役割を担う。このフィルターの重みが学習対象である。pooling 層は、convolution 層から出力された特徴量に対してサブサンプリングを行うことで次元を下げ、かつ局所的な入力の変化に対して鈍感にしている。一般に、領域内で最大値を出力する max pooling や、平均値を出力する average pooling が用いられる。CNN では、convolution 層と pooling 層を交互に経ることによって非線形な特徴量抽出を行った後、全ユニットが結合している層によってそれらの特徴量を統合し、分類を行う。

4. 実験

4.1 傾向スコア推定

ロジスティック回帰、GBDT、CNN を用いて傾向スコアの推定を行った。傾向スコアの学習は、観測された共変量の集合を入力として、処理の割り付けを表すダミー変数 `treat` を教師信号とした教師あり学習によって行った。これによって推定された値は、処理が割り付けられた状態への所属確率として得られ、正規化された確率値となる。

本稿で扱う MAR の問題設定では、目的変数を *target* とおくと、変数 *treat* が説明変数となり、観測された共変量がこれら二つの変数と交絡していると見なせる。よって、傾向スコアを推定する際には、目的変数を入力に含めないようにしなければならない。

`lalonde`、`lindner`、`ACTG175` の 3 つのデータセットを対象に計算機実験を行った。`lalonde` 及び `ACTG175` データセットでは、*treat* は `treat` という変数として与えられており、`lindner` データセットでは `abcix` が *treat* にあたる。また、`lalonde` では `re78` が、`lindner` では `lifepres` が *target*

変数にあたる。ACTG175 では、*target* 変数 cd496 が欠測しているためこの欠測指標を新たに missing と定義し、これを *treat* とした。

各モデルのハイパーパラメータは、交差検証を 5 fold で行い傾向スコアの推定精度が最も高くなったものを選択した。特に CNN では、convolution 層のフィルターサイズは 3×1 とし、pooling 層では 2×1 の max pooling を行った。活性化関数は ReLU を用い、バッチサイズを 50 として 20 epoch 学習させた。

4.2 傾向スコアによるマッチング

推定された傾向スコアの値を用いて貪欲マッチングを行った。傾向スコアはロジスティック回帰、GBDT、CNN によって推定されたものをそれぞれ用いた。欠測群及び観測群からマッチングされたペアの共変量が、ランダムマッチングと比べてどれだけ良くなっているかを評価するために、以下の二つの方法を用いた.

4.2.1 t 値による評価

マッチングを取った欠測群と観測群について t 値を計算し、各学習器について比較することで評価した。ここで、欠測指標 $r \in \{0, 1\}$ を用いて、マッチングを取ることで得られた欠測群と観測群両群の共変量ベクトルの集合を $\boldsymbol{x}_r$ と表し、それらの平均を $\overline{\boldsymbol{x}}_r$、分散を $\bar{s}_r$ とおくと、t 値とは、

$$t = \frac{\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}} \tag{4}$$

と表される量であり、分布の近さを分布の平均と分散を用いて表現している。t 値が小さいほど、分布として似ていることを示す指標である。

4.2.2 Kullback-Leibler 情報量による評価

マッチングを取った両群に対して、各共変量について Kullback-Leibler (KL) 情報量を計算、各学習器について比較することで評価した。連続確率分布 P, Q に対して、$D_{\mathrm{KL}}(P||Q)$ は P の Q に対する KL 情報量を表し、p, q をそれぞれ P, Q の確率密度関数とおくと、

$$D_{\mathrm{KL}}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{5}$$

と書ける。本稿では、P, Q をそれぞれ欠測群, 観測群のある共変量の連続確率分布とおくことによって、KL 情報量が小さいほどその共変量の分布が二群で似通っていることを示す指標とした。

4.2.3 シミュレーション

まず、データを 5 分割し、それらのうち 4 つをトレーニングデータ、残る 1 つをテストデータとするようなデータ分割を、分割の方法をランダムに変化させながら 10 回行うことで新たなデータセットを作成した。次に、トレーニングデータのうち 1 つを用いて各学習器で傾向スコア $p(r = 1|\boldsymbol{x})$ を学習させた。さらに対応するテストデータで傾向スコアを算出し、それに基いて貪欲マッチングを行い、t 値と KL 情報量を計算した。以上を 5 分割したデータそれぞれについて行うことを 10 回繰り返し、得られた全ての t 値を lalonde, lindner, ACTG175 それぞれのデータセットについて学習器ごとにプロットしたものが Figure 1, 2, 3である。

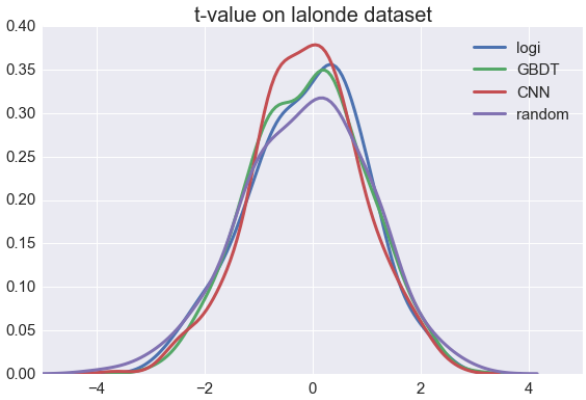

Figure 1: lalonde

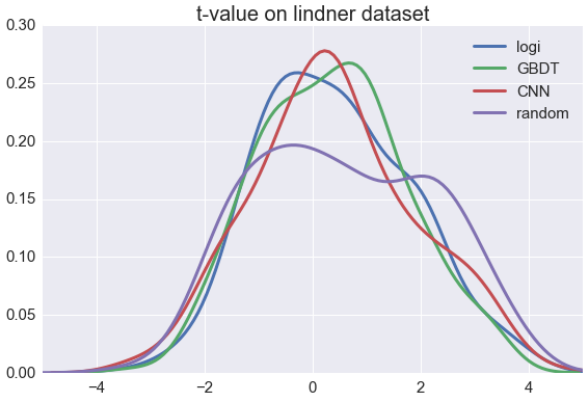

Figure 2: lindner

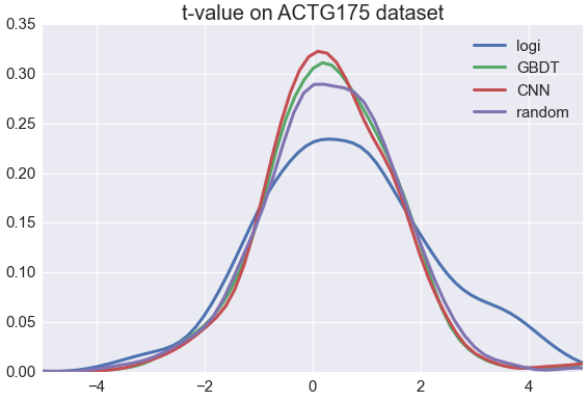

Figure 3: ACTG175

これらの図から、ロジスティック回帰に比べて GBDT や CNN のような非線形な手法の方がより t 値が 0 に近いところに分布しており、特に CNN が高い性能を示していること

がわかる。

また、KL 情報量についても同様にして各データセット、学習器ごとに各共変量の KL 情報量を計算した。共変量の次元が大きいため、CNN において大きな改善が見られた共変量と改善が見られなかった共変量をそれぞれ順に 2 つずつ Table 1, 2, 3に示した。

Table 1: lalonde での各共変量の KL 情報量

	変化大		変化小	
手法	u75	re74	married	black
ベースライン	6.534	25.53	16.18	3.530
LogisticR	4.042	20.43	15.67	3.567
GBDT	4.868	18.78	14.66	3.212
CNN	0.589	6.88	15.48	3.142

Table 2: lindner での各共変量の KL 情報量

	変化大		変化小	
手法	cardbill	height	stent	acutemi
ベースライン	0.468	0.00583	8.00	16.76
LogisticR	0.184	0.00534	7.90	16.30
GBDT	0.446	0.00563	8.11	15.88
CNN	0.127	0.00530	7.88	16.17

Table 3: ACTG175 での各共変量の KL 情報量

	変化大		変化小	
手法	pidnum	cd80	karnof	wtkg
ベースライン	1.263	0.302	0.00594	0.0445
LogisticR	1.209	0.278	0.00534	0.0440
GBDT	1.219	0.298	0.00583	0.00439
CNN	1.005	0.270	0.00593	0.00438

CNN での改善率が大きかった共変量については、いずれのデータセットにおいても他の学習器やベースラインを凌ぐ結果が得られた。また、CNN での改善率が小さかった共変量についても、他手法と並ぶか僅かに劣る程度であった。

以上のような結果から、傾向スコア推定における非線形な手法、特に CNN の有効性が確認できる。このような結果になった理由は大きく分けて二つあると考えられる。一つは、欠測に関わる変数の任意性である。MAR の仮定から、欠測が生じるか否かは共変量から説明できるが、欠測の生じるパターンは一意ではない可能性がある。ロジスティック回帰のようなモデルでは、共変量ごとに重みが決定されてしまうので、複数の欠測パターンに対応することは難しい。しかし、CNN では重みはフィルターとして学習されるため、複数の欠測パターンにも対応可能である。もう一つは max pooling による強力な変数選択である。欠測のメカニズムが不明である中で MAR の仮定を満たすために、共変量は多次元になることが多い [9] ので、欠測に関わる共変量は全体の一部である可能性がある。そのため、モデル自体が変数選択可能なものが望ましいが、CNN では max pooling によるサブサンプリングによってそれが自然と行われている。

5. 結論

傾向スコア推定によく用いられる複数のデータセット全てにおいて、従来用いられてきたロジスティック回帰等の線形モデルを GBDT や CNN のような非線形なモデルに変更して傾向スコアを推定することによってマッチング精度の向上が見られた。特に CNN での改善が大きく、これは、欠測に関わる共変量の組み合わせが複数パターン存在する可能性があり、かつ有効な変数は全体の中でも一部である、というスパース性を満たしていることが要因であると考えられる。今後は、推定された傾向スコアに基づくマッチングによって欠測推定を行うことを目標としたい。

謝辞

本研究は JSPS 科研費 26730130、15K12112 の助成を受けた。

参考文献

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

[2] J. Carpenter and M. Kenward. *Multiple Imputation and its Applications*. Wiley, 2013.

[3] S Guo and M. W. Fraser. *Propensity Score Analysis: Statistical Methods and Applications, Second Edition*. Sage, 2015.

[4] Y. LeCun, L. Bottou, Y. Benjio, and P. Haffner. Gradient-based learning applied to document recognition. 86(11):2278–2324, 1998.

[5] T. Lei, R. Barzilay, and T. Jaakkola. Molding cnns for text: non-linear, non-consecutive convolutions. 2015.

[6] Roderick J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd edition, 2002.

[7] P. R. Rosenbaum. *Design Of Observational Studies*. Springer, 2010.

[8] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. 70(1):41–55, 1983.

[9] D. B. Rubin. *Matched Sampling for Causal Effects*. Cambridge : Cambridge University Press, 2006.

[10] T. N. Sainath, R. J. Weiss, K. W. Wilson, and O. Vinyals. Learning the speech front-end with raw wave from cldnns. 2015.