論文 / 著書情報
Article / Book Information

論題(和文)	系統的な欠損を持つデータにおける項目序列決定手法の提案
Title(English)	Proposal of item ordering method for data with systematic loss
著者(和文)	青木亮麿, 北澤正樹, 高橋聡, 吉川厚, 山村雅幸
Authors(English)	Ryoma Aoki, Masaki Kitazawa, Satoshi Takahashi, Atsushi Yoshikawa, Masayuki Yamamura
出典(和文)	教育システム情報学会 2019年度学生研究発表会, , , pp. 63-64
Citation(English)	, , , pp. 63-64
発行日/Pub. date	2020, 3

系統的な欠損を持つデータにおける項目序列決定手法の提案

Proposal of item ordering method for data with systematic loss

青木 亮磨[*1], 北澤 正樹[*1], 高橋 聡[*2], 吉川 厚[*1], 山村 雅幸[*1]
Ryoma AOKI[*1],Masaki KITAZAWA[*1],Satoshi TAKAHASHI[*2],Atsushi YOSHIKAWA[*1],Masayuki YAMAMURA[*1]
[*1] 東京工業大学情報理工学院
[*1]School of Computing, Tokyo Institute of Technology
[*2] 関東学院大学理工学部
[*2]School of Science and Engineering, Kanto Gakuin University
Email: aoki.r.ae@m.titech.ac.jp

あらまし: 教育においてデータを活用して教育施策に活用する動きが近年盛んになっており，多くの教育に関するデータが雑誌や本，WEB 上で公開されている．本研究では大学入試に関するデータを用いた大学の入試難易度序列の決定手法を提案する. 各高校の公開している大学合格実績データは高校の実力と大学入試の難易度に依存する系統的な欠損を持ち, 高校がターゲットとする大学のみ合格率が高いという性質を持つ．ここで，高校からの合格率の高い大学の入試難易度が似ていると考えると，合格率の高い部分にのみ注目することで系統的な欠損のあるデータで入試難易度を比較できる. 得られた序列を塾が公開する偏差値による序列との順位相関で評価した結果, 0.87~0.89 となり, 従来手法より良い結果が得られた.
キーワード： Evidence Based Education, 序列決定, 系統的欠損, 公開データ

1. 背景

近年，教育分野でも Evidence Based Education(以下 EBE)として，教育は証拠に基づいているべきであるという考えが重視されるようになってきた(1)．エビデンスとして被験者個人のデータを収集・分析した量的研究が多い(2)．しかし，教育効果が異なると想定される教育手法をデータ取得のために対象群に適用する場合に教育倫理として問題となることがあり，個人データの入手には多大な困難がある．

一方，教育に関わるデータは web 上や本などにおいて公開されており，膨大な量がある．公開されているデータは個人データではなく集計データであり，個人の学習などのデータとして使用することは難しいが，学校などの組織単位でのエビデンスとして用いることは十分可能であると考えられる．しかし，この時に系統的な欠損が問題になることがある．

系統的な欠損とは，受験者が全ての問題を解かないことから生じる，集計データにおける理想的な状態のデータと実際の状態のデータの差である．例えば，高校の大学合格実績データについて考える．大学入試合格実績として理想的なデータ状態は全ての受験生が全ての大学を受験し，難易度が高い大学にはわずかしか合格せず，難易度が下がるにつれて合格者数が増えていくというものである．しかし，入試難易度が低い大学には受験をしないので合格者が多くならない．これは金銭面的，時間的制約により生じる現象である．この状況を図 1 に示す．

従来の序列を決定する手法では，対象の直接比較を行う手法(3)かアンカー項目を用いる手法(4)が用いている場合が多い．しかし，系統的な欠損のあるデータとして高校の大学合格実績データについて考えた時に，大学が直接対決しているわけでも，アンカー項目があるわけでもないため，従来の手法を適用することは難しい．

そこで，本研究では大学入試合格実績データを用いて，系統的欠損データに対応した序列決定手法を作成することを目的とする．

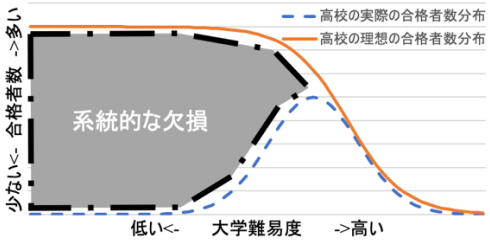

図 1 大学入試における系統的な欠損

2. 手法

2.1 データ

各高校が公開している合格実績と卒業者数，複数の会社が公開している大学偏差値データを使用する．今回は地域性を考慮するため，東京都，千葉県，埼玉県，神奈川県にある大学/高校のみのデータを用いる．高校は 886 校，大学は 75 校となった．

2.2 評価方法

大学入試難易度として，塾の公表する偏差値がある．例えば，3 社が公開している偏差値から得られる大学序列の順位相関をとると表 1 のようになり，各社の偏差値序列の順位相関は 0.98 と高い. そこで，大学入試難易度序列の評価は，各偏差値から得られる序列とのスピアマンの順位相関を取るものとする．

表 1 各社の偏差値序列の順位相関

	B 社	C 社
A 社	0.98	0.98
B 社		0.98

2.3 従来の手法への適用

従来の序列決定手法の例として直接比較を行う手法として Elo Rating[3], アンカー項目を用いる手法として IIHF Ranking[4]に合格実績データを適用する．得られた序列を各社の偏差値序列と順位相関を取ったものを，表 2 に示す．系統的欠損データから従来手法での序列決定が困難なことが確かめられた．

表 2 従来手法から得られる序列の評価

	A 社	B 社	C 社
Elo Rating	0.21	0.22	0.20
IIHF Ranking	-0.05	-0.07	-0.07

2.4 提案手法

各高校からの各大学への合格率を用いて序列を決定する．大学と高校を交互に難易度/実力の高い順番に並べていく手法である．作成した手法のアルゴリズムは以下のとおりである．

1. 大学群U(初期値は東京大学と東京医科歯科大学)を順位 i（ループ回数）とする
2. 大学群Uへの合格率の合計が高い高校を$|U| * m/n$だけ選択し，これを高校群 H とする
3. 高校群Hからの合格率を用いて，順位未決定の大学を X-means を用いてクラスタリング
4. 高校群Hからの合格率の平均が最も高いクラスタに所属する大学を新たな大学群Uとする
5. 順位未決定の大学が無くなった場合終了．順位未決定の大学が存在する場合は 1 へ

このアルゴリズムでは試行ごとに結果が異なるため，シミュレーションを複数回行い，各大学の平均順位をとる．評価は以下の式を用いて行う．

$$s_i = \sum_j^N r_{i,j} \qquad (1)$$

ここで，s_iを大学iの評価値，$r_{i,j} \in \mathbb{N}$をシミュレーションjにおける大学iの順位，Nをシミュレーション回数とする．求めたスコアを昇順に並べたものを大学の序列とする．

3. 結果

式(1)において$N = 4000$とし，作成された大学序列を偏差値から得られる大学序列とスピアマンの順位相関を取ることで評価を行う．各社の偏差値との順位相関を表 3 に示す．$N = 4000$の時の序列と各社の偏差値序列との順位相関は 0.87~0.89 となった．また，$N = 4000$までの順位平均から得られる序列との順位相関の推移を図 2 に示す．

図 2 より，シミュレーションを重ねていくことで序列の安定化と順位相関係数の向上を図ることができた．受験生は全ての大学を受験できないため実力に合った大学のみを受験するという系統的な欠損と，高校には似た実力者が集まることから，各高校の合格実績を大学入試難易度を軸とした合格者数の度数分布に表した時にピーク(山)が生じる．このピークでピークシフトを行なっていくことで序列を決定する手法を作成した．各社の偏差値序列の平均順位より 10 位以上低くなった大学として武蔵大学, 立正大学，獨協大学，創価大学などがある．これらの大学は系列高校からの合格率が 100%に近く，系列高校の生徒は他の大学を受験しないため，ピークが尖ってしまう．このため，順位が低くなったと考えられる．逆に，各社の偏差値序列の平均順位より 10 位以上高くなった大学として東京農業大学，電気通信大学，東京芸術大学などがある．これらの大学は単科大学であり専門性が高いため，中上位高校において一番人気というわけではないが高い人気と合格者数を誇る大学である．このため，順位が高くなったと考えられる．

表 3 作成された序列と偏差値序列との順位相関

	A 社	B 社	C 社
$N = 4000$	0.89	0.87	0.88

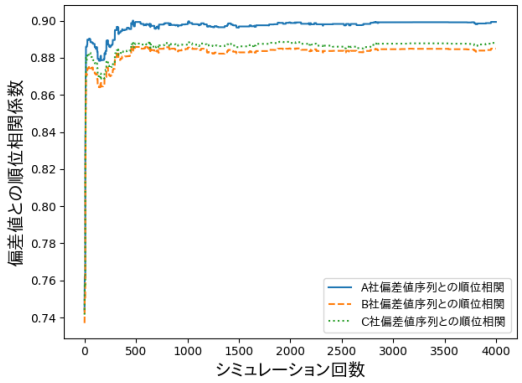

図 2 $N = 4000$までの順位相関の推移

4. おわりに

本研究では，公開されている教育データを利活用する際に問題になることがある系統的な欠損という問題に対して，序列決定手法を作成した．作成した手法では偏差値序列との順位相関が 0.87~0.89 となり，従来手法のものより良い結果が得られた．

参考文献

(1) Alexander W. Wiseman, “The Uses of Evidence for Educational Policymaking: Global Contexts and International Trends”, Review of Research in Education, Vol.34, No.1, pp.1-24(2010)
(2) 耳塚寛明, “小学校学力格差に挑む だれが学力を獲得するのか”, 教育社会学研究, 第 80 号, pp23-39(2007)
(3) Arpad E. Elo. “The Rating Of Chess Players Past & Present.” Arco Publishing, (1978)
(4) International Ice Hockey Federation. https://www.iihf.com/en/worldranking. (参照 2020-01-21)