1ZH-01

不良回答を含むアンケートデータの信頼性向上手法について

村木鴻介† 髙橋朋治† 小田陽平† 丹生谷英子† 須子統太†

早稲田大学社会科学部†

1. はじめに

近年,様々な場面で Web を利用したアンケート収集が行われている.Web アンケートは手ごろに行えるという利点がある反面,回答者が正しく設問に回答しない,いわゆる不良回答が混入しやすいという問題がある.不良回答のパターンとしては,選ばれる選択肢が中心付近の狭い範囲に限定されるパターン,同一の回答のみを選択するパターン,無回答が多くなるパターンなどが知られている.[1]

こうした不良回答に対して,従来は上記のような回答パターンに当てはまる回答を不良回答として規定し,除去することでデータのクリーニングが行われている．また，あらかじめアンケートに回答者の矛盾を検出する設問を組み込んでおき,矛盾した回答をする回答者を除去するという方法が取られている.しかし,一つ目の手法では正しく回答した結果と規定した不良回答パターンが一致してしまった場合，本来正しく回答したはずのデータが除去されてしまうという問題がある.また,二つ目の手法はアンケートを実施する前に設問を設計する必要があり,すでに得られているアンケートデータの不良回答を除去する場合には用いることができない.

本研究では上記とは別のアプローチとして,ランダム回答法と呼ばれる手法を応用することで,不良回答を除去することなく，直接的に回答分布を推定する方法を提案する．そのもとで,実際の規模アンケートデータを用いた提案手法の有効性の検証を行う.

2. ランダム回答法

本研究に応用したランダム回答法について説明する.ランダム回答法とはアンケート調査により回答分布を推定するための手法である.これは個人として公にしたくない内容に関するアンケート調査の場合,回答者が正確に回答するとは限らず,正確な回答分布の推定を行うことができない場合などに用いられる.

このような状況に対し,Warner は回答の「Yes」と「No」が反転するような質問を用意し,回答者に質問自体を確率的に選択させるという方法を提案した.これは回答に意図的にノイズを加えているとみなせる．Warner はこのようなノイズを含むデータを用いて回答分布を推定する手法を提案している.[2]

3. 提案手法

3.1 不良回答のモデル化

前述のランダム回答法と同様に，不良回答をアンケートに加えられたノイズとしてモデル化する.いま，設問数がD個のアンケートに対し,d番目の設問に対する回答を$a_d \in \{0,1,\cdots,k-1\}$とする.ここで$k$は回答の選択肢の数とする.$D$個の設問全てに対する回答パターンを（$a_1, a_2, \cdots a_D$）と表す.真の回答パターンを$X$とした時に,不良回答によりノイズの乗った回答パターンをYで表す.不良回答の発生確率を，条件付確率$\mathrm{P}(Y|X)$でモデル化する事とする.

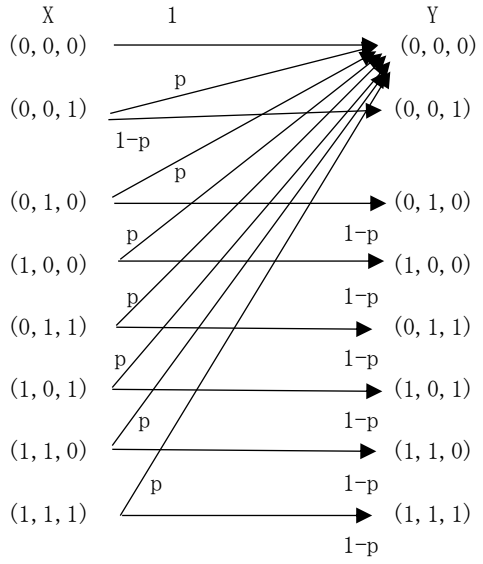

図１:不良回答モデルの例

A Reliability Improvement Method for Questionnaire Date with Defective Answers

Kohsuke Muraki, Tomoharu Takahashi, Youhei Oda, Eiko Nyunoya, Tota Suko Waseda University Social Science Study

例えば$D=3$で$k=2$とし，条件付き確率として上記の図1のような遷移確率を仮定する．これは，ある確率pで不良回答をする回答者が出現し，不良回答者は実際の回答によらず一律に同じ回答（0, 0, 0）をするというモデルである．

これ以外にも遷移確率を変えることで様々な場合を表現できる．

3.2 回答分布推定法

提案する回答分布推定方は，様々な遷移確率に対し構成可能であるが，以下簡単のため図1で示したモデルを例に説明する．

図1のモデルの場合，回答パターンYの確率は次のように表される．

$$P(Y=000) = P(X=000) + p\times\{P(X=001)+\cdots+P(X=111)\} \quad (1)$$

$$P(Y=001) = (1-p)\times P(X=001) \quad (2)$$

$$\vdots$$

$$P(Y=111) = (1-p)\times P(X=111) \quad (3)$$

ここで，得られたサンプルから推定した$\mathrm{P}(Y)$の推定値$\hat{P}(Y)$を用いて，$\mathrm{P}(X)$を以下で推定する．

$$\hat{P}(X=001) = \frac{\hat{P}(Y=001)}{1-p} \quad (4)$$

$$\vdots$$

$$\hat{P}(X=111) = \frac{\hat{P}(Y=111)}{1-p} \quad (5)$$

$$\hat{P}(X=000) = \hat{P}(Y=000) - p\{\hat{P}(X=001)+\cdots\hat{P}(X=111)\} \quad (6)$$

4. 評価実験

実際のアンケートデータに，仮定したモデルに従い，人工的に不良回答を混入させたデータを作成することで，前述の手法の有効性を検証した．使用するデータは，株式会社アサツーディ・ケイ生活意識調査2017，消費意識に関する4択（選択肢は0, 1, 2, 3）アンケートのデータの中の5項目を用いる．データ総数は15042個であった．

実験では図2で示すモデルに従い不良回答を発生させた．これは，不良回答者が一律に(1, 1, 1, 1, 1)を回答するモデルで，$p=0.05$としたもとで確率的に不良回答を発生させた．

このもとで，不良回答を混入する前の正しいアンケート結果から集計した回答分布と，以下の3つの方法で導出した回答分布との分布の距離を測定した．一つは不良回答を混入させたデータをそのまま集計した回答分布，もう一つは(1, 1, 1, 1, 1)の回答パターンを全て除去したデータから集計した回答分布，最後に提案手法で推定した回答分布である．分布間の距離は KL 距離で算出した．

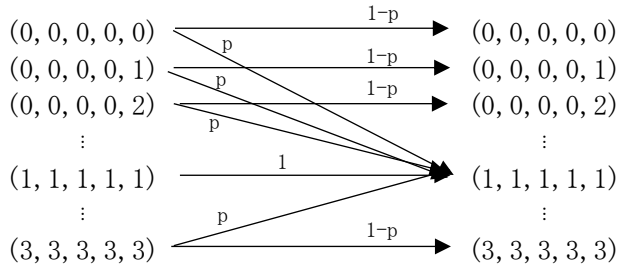

図2：実験で使用した不良回答モデル

実験結果を表1に示す．提案手法は単純に不良回答パターン(1, 1, 1, 1, 1)を削除した推定法よりも，正しい回答分布との距離が近くなっており，良い推定法になっていることがわかる．

表1：各項目の正解データとの KL 距離

	不良回答混入	(1, 1, 1, 1, 1)削除	提案手法
Q1	0.001325	0.000234	0.000021196
Q2	0.004343	0.000747	0.000053671
Q3	0.00128	0.00023	0.000045262
Q4	0.000759	0.000108	0.000038266
Q5	0.002722	0.000512	0.000015142

5. まとめ

本研究では，不良回答を含むアンケートデータからの回答分布推定問題に対し，確率的な不良回答モデルを利用した回答分布推定手法の提案を行った．そのもとで，実際のアンケートデータに対して人工的に不良回答を付加することで，提案手法の有効性を検討した．

しかし，提案した手法は不良回答モデルを示す遷移確率が既知である場合にしか適用できない．遷移確率が未知である場合の分布推定法については今後の課題としたい．

参考文献

[1] 山田文康ほか，”アンケートにおける「不良回答」の回答特性と分析結果に与える影響に関する研究，”日本社会情報学会 25 回全国大会，2010.

[2] Warner S. L., “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” journal of the American Statistical Association, Vol.60, No.309, pp.63-69, 1965.