The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Bayesian Networkを用いた動的アンケートシステムによる設問の回答推定

Estimation of the answers to unanswered questions by Adaptive questionnaire system based on Bayesian network

田村　脩　　櫻井　瑛一　　本村　陽一
Shu Tamura　　Eiichi Sakurai　　Youichi Motomura

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

One of the most effective ways to estimate user types is to administer a questionnaire. In most cases, a questionnaire has multiple questions, and requires considerable time to complete.To make the process time-efficient, we proposed an adaptive questionnaire system based on a Bayesian network in our previous study. Under this system, a question evaluation function is used to ensure that subsequent questions are adapted to the respondents' answers to preceding questions. Afterward, the respondents are categorized according to user types. In this study, we modify the question evaluation function in the previous system to enable it to estimate the answers to unanswered questions based on the estimated user type. Experimental results show that although the proposed method is inferior to the method that directly estimates answers without using segments, it is superior to the original method that is specialized for user type estimation.

1. はじめに

昨今，データを集めてユーザーの客観的なモデル化を行うために，ユーザーの特徴を把握するアンケートが行われる機会が増加している．例としては，顧客のアンケートデータにProbabilistic Latent Semantic Analysis[Hofmann 99] (以降 PLSAと略す) を行うことで，ユーザのセグメントを得たのち，アンケートの設問とセグメントに基づくベイジアンネットワークを作成することで，セグメントの特徴を見出した研究が存在する[本村 16].

上の例は，すでに行われたアンケートのデータから，ユーザーのモデル化が行われている．一方で，未知のユーザーに対しては，各ユーザーの回答結果からどのセグメントに属するか等を判定することが重要になる．しかし，未知のユーザーに対してアンケートにすべて回答してもらうことは，ユーザー側の負担が非常に大きいため，回答精度や回答率が落ちる恐れがある．よって，少量の設問に回答してもらうことでユーザーの所属するセグメントを推定する必要があり，ユーザーに質問する設問の選定が非常に重要である．また，アンケートは複数の設問から成り，各設問は問題文と複数の選択肢で構成されている．アンケートの構造上，設問から問題文と選択肢を切り離すことは好ましくないため，設問単位での選定が必要になってくる．

以上の要因から我々は，ベイジアンネットワークによるセグメント説明モデルを用いて，有力な設問を各ユーザーの回答結果に応じて逐次提示する動的アンケートシステムを過去に提案した [田村 18]．その結果，設問をランダムに提示した場合と比較し，より少ない設問数で同程度のセグメント推定精度を得ることに成功した．

本論文では，より実践的なタスクとして，限られた設問の回答から未出題の設問に対する回答を推定することを考える．本タスクにおいては，セグメントの代替として推定する設問(以下,「推定設問」とする) の変化量を算出することで動的アンケートシステムは適用可能である．しかし，システムによる設問選択の計算量が推定設問の選択肢数に依存するため，社会実装が困難であるという問題点が存在した．そこで，設問選択の計算量を推定設問の選択肢数に依存することなく，推定設問回答の推定精度をより高く保つことのできる設問選択アルゴリズムを新たに提案する．提案手法の有用性を判断するため，アンケートデータを用いて実験及び評価を行った．

連絡先: 櫻井　瑛一，国立研究開発法人産業技術総合研究所 人工知能研究センター，東京都江東区青海 2-4-7，03-3599-8916，e.sakurai@aist.go.jp

2. 動的アンケートシステム

先に，動的アンケートシステム内で使用される手法を紹介し，続いて我々が過去に提案した動的アンケートシステムについて説明する [田村 18]。

2.1 PLSA

PLSA(Probabilistic Latent Semantic Analysis) は文書分類のために考案されたモデルである [Hofmann 99].

このモデルは，文書に出現する単語 w は話題 k によって変化し，各文書 d は話題 k の混合によって出来上がっていると考える．この話題 k がいわゆる潜在クラスである．このモデルでは文章，単語，潜在クラスの同時確率を次のように定義する:

$$P(w,d,k) = \sum_k P(w|k)P(d|k)P(k).$$

この同時確率から文章データセット $\mathcal{D}$ に対する尤度を計算し，EM 法を用いて尤度最大化されるように，$P(w|k), P(d|k), P(k)$ を推定する．

今回の分析においては，ユーザ d とそのアンケートの回答 w との間にセグメント k（性質）が存在すると考え，PLSA を行い，各ユーザの所属するセグメントを得た．

2.2 Bayesian Network

ベイジアンネットワークとは，条件付き確率により，データ間の潜在的な依存構造をモデル化する手法である．例えば，X_1, X_2, X_3, X_4 の 4 個の変数が存在したときに，その同時確率を

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_4)P(X_4)$$

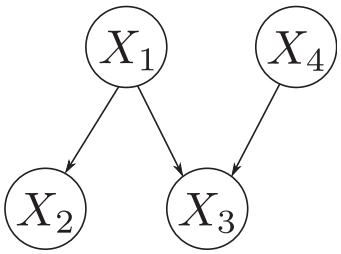

図 1: ベイジアンネットワークの例

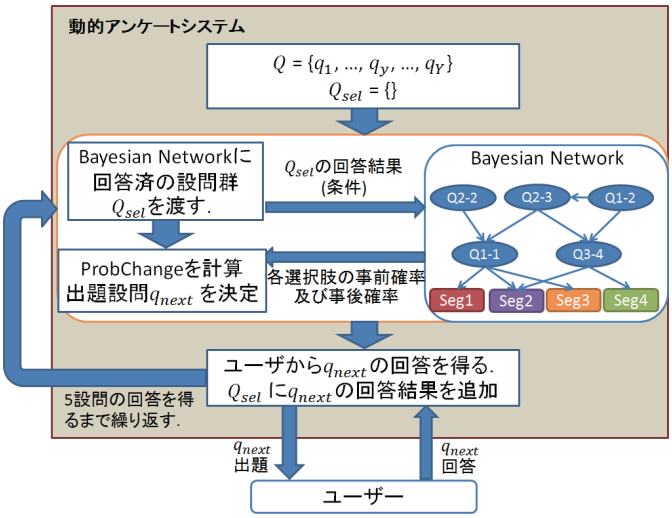

図 2: 動的アンケートシステムの外観図

と表現できたとすると，その関係をグラフとして図 1 のように表現する．このような条件付き確率に基づいたモデル化の手段をベイジアンネットワークという．データからこのネットワーク構造を推定する場合，各変数が条件付き確率としてもっとも結びつきが強いものから選択され，全体の構造の複雑さとのトレードオフによりグラフ構造が推定されることになる．

今回の分析では，各変数 X_i がアンケートの設問の選択肢及び，セグメントを示す．

2.3 動的アンケートシステム

[田村 18] にて提案した動的アンケートシステムについて説明する．まずアンケートデータそのものに対して PLSA を用いてクラスタリングを行い，得られたセグメント及びそれを説明する設問に関するベイジアンネットワークを構築する．その後，セグメントにエビデンスを与えた際の各選択肢の選択確率の変化量を算出する．そして，変化量の合計値 (ProbChange) が大きい設問を「重要な設問」であるとし，次に出題する設問とする．なお，それまでの回答結果についてはすべてエビデンスとして加えた状態にするため，各ユーザーの回答結果に応じて逐次的に異なる設問が出題される．動的アンケートシステムの概観図を図 2 に示す．

図 2 において，Q は設問の集合，Q_{sel} は既に回答された設問の集合を表す．$ProbChange$ は各設問単位での確率値の変化量を表す．

2.3.1 設問追加時の評価値 $ProbChange$

セグメントを設問の各選択肢をが説明する構造のベイジアンネットを考える．セグメントの集合を $C = \{c_1, ..., c_x, ..., c_X\}$，設問の集合を $Q = \{q_1, ..., q_y, ..., q_Y\}$，設問の集合 Q における y 番目の設問 q_y が持つ選択肢の集合を $q_y = \{s_1, ..., s_k, ..., s_{K_y}\}$ とし，既に回答された設問集合を Q_{sel} としたとき，新たなエビデンスを付与した際の設問 q_y の確率値の変化量 $ProbChange(q_y, Q_{sel})$ を以下の式で定義する．

$$ProbChange(q_y, Q_{sel}) = \sum_{k=1}^{K_y} d(s_k)$$

$$d(s_k) = \sum_{x=1}^{X} \sum_{z\in\{0,1\}} \sum_{i}^{I} P(s_k|Q_{sel}, c_x = z)_i \log \frac{P(s_k|Q_{sel}, c_x = z)_i}{P(s_k|Q_{sel})_i}$$

ここで，$P(s_k|Q_{sel})$ は，既に回答された設問 Q_{sel} の回答をエビデンスとして与えた際に推測される選択肢 s_k の選択確率分布である．そして，$P(s_k|Q_{sel}, c_x = z)$ は，既に回答された設問 Q_{sel} の回答及び，セグメント c_x に z(0 または 1) のエビデンスを与えた際に推測される選択肢 s_k の選択確率分布である．ここで，$P(s_k|Q_{sel})$ と $P(s_k|Q_{sel}, c_x = z)$ の KL ダイバージェンスをとり，各 z，全セグメントについて和をとることで，選択肢 s_k の選択確率の変化量とした．なお，$c_x = 0$ は，セグメント c_x に属さないという意味であり，$c_x = 1$ は，セグメント c_x に属するという意味である．さらに，この値を設問 q_y 中の全選択肢で合計することで，設問 q_y の変化量 $ProbChange(q_y, Q_{sel})$ とした．

2.3.2 逐次的設問選定の方法

各セグメントの所属確率の変化に敏感に反応する選択確率を持つ選択肢を含む設問が，優先して選定するべき重要な設問であると考えられる．そこで，$ProbChange$ が最も大きい設問を，次の設問として採用する．この方法によって，セグメントへの影響が大きい設問を優先的に採用できると考えられる．以下に具体的な式を示す．

既に回答された設問集合を Q_{sel} とすると，アンケートシステムで次に質問する設問 q_{next} は，以下の式によって決定される．

$$next = \underset{y|1\le y\le Y, y\notin Q_{sel}}{\arg\max} \quad ProbChange(q_y, Q_{sel})$$

つまり，まだ質問の終了していない設問群のうち，$ProbChange$ の値が最も大きい設問を採用する．

3. 設問の回答推定に特化した設問選択アルゴリズムの提案

本項では，設問の回答推定に特化した設問選択アルゴリズムを提案する．能動的に設問の回答を推定するために，$ProbChange$ の定義を以下に拡張した．

推定設問を $q_{esti} = \{e_1, ..., e_k, ..., e_{K_{esti}}\}$ とする．そして，q_{esti} を除いた設問の集合 Q における y 番目の設問 q_y が，選択肢の集合 $q_y = \{s_1, ..., s_k, ..., s_{K_y}\}$ を持ち，既に回答された設問集合を Q_{sel} としたとき，新たなエビデンスを付与した際の設問 q_y の確率値の重み付変化量 $W\_ProbChange(q_y, Q_{sel})$ を以下の式で定義する．

$$W\_ProbChange(q_y, Q_{sel}) = \sum_{k=1}^{K_y} d(s_k)$$

$$d(s_k) = \sum_{x=1}^{X} \sum_{z\in\{0,1\}} w_{(x,z)} \sum_{i}^{I} P(s_k|Q_{sel}, c_x = z)_i \log \frac{P(s_k|Q_{sel}, c_x = z)_i}{P(s_k|Q_{sel})_i}$$

$$w_{(x,z)} = \sum_{k=1}^{K_{esti}} \sum_{i}^{I} P(e_k|Q_{sel}, c_x = z)_i \log \frac{P(e_k|Q_{sel}, c_x = z)_i}{P(e_k|Q_{sel})_i}$$

$W\_ProbChange(q_y, Q_{sel})$ が $ProbChange(q_y, Q_{sel})$ と比較して異なる点は，各セグメントについて重み $w_{(x,z)}$ を定義し，重み付合計を選択肢 s_k の変化量としている部分である．各セグメントについて算出される重み $w_{(x,z)}$ については，各セグメントにエビデンスを与えた際の，推定設問の全選択肢の変化量の合計である．推定設問を推定する際に重要なセグメントというのは，推定設問に大きな影響を与えるセグメントであるため，そのようなセグメントに重きを置いて $ProbChange$ を算出するという考え方である．

4. 実験

今回の実験では，動的アンケートの種類・ベイジアンネットモデルの種類を変化させて実験を行った．

4.1 使用したデータセット

アンケート対象として 18 歳以上で車を所有し，車購入の決定権者でかつ 5 年以内の購入者の中で，車に対して安さのみを求めない人を対象者にインターネット調査を行った結果のデータを今回は使用した．設問の内容は，回答者のデモグラフィック属性と，車の購入に対する重視点，車に対するわくわく感などの感性的な質問である．このデータでは，総数として 4164 名のデータがあり，その中でも，回答に対する誠実さを問う設問に正確に答えた 3373 名のデータを使用した．

4.2 実験概要

本実験の概要について述べる．

まず，実験を行うにあたり，3373 名の全回答について，PLSA により 7 つのセグメントに分類した．続いて，各設問の選択肢及びセグメントを用いて，ベイジアンネットワークを作成した．なお，ベイジアンネットモデルの構築には，ベイジアンネットワーク構築ソフトウェア Bayonet[本村 03] を用いた．

本実験では，推定設問の回答を推定するという目的のため，最終的に推定設問の選択確率を出力し，実際の回答結果と比較した．本研究で設定した推定設問は「車を購入する時に重要視するポイントは何ですか？」という複数回答設問であり，選択肢には「小回りが利く」「燃費性能」などを含む．選択肢数は 21 である．

なお，評価に使用するデータセットは，PLSA 及びベイジアンネットワークを作成した際と同様のデータを使用した．また，実験条件について以下に述べる．本研究ではベイジアンネットワークのモデル・動的アンケートシステム上で与えるエビデンス情報・設問選択の方法をそれぞれ変化させて比較実験を行った．以下の項で実験条件の詳細について解説する．本論文内では，実験条件を (ベイジアンネットワークのモデル)-(与えるエビデンス情報)-(設問の選択方法) と表記する．

4.2.1 ベイジアンネットワークのモデル

以下の 2 種類のベイジアンネットモデルを作成した．

- Q&C: 設問選択肢とセグメントから成るネットワーク．推定設問選択肢をセグメント及び各設問選択肢が説明し，各セグメントを各設問選択肢が説明する．
- Qonly: 設問選択肢のみから成るネットワーク．セグメントを用いず，推定設問選択肢を他の設問選択肢が説明する．

4.2.2 動的アンケートで与えるエビデンス情報

与えるエビデンス情報も，2 通り用意した．

- Qesti: 推定設問
- C: セグメント（ベイジアンネットのモデルが Q&C の場合のみ使用可能）

q_esti については，計算量が推定設問の設問数に依存するため，本実験の様に設問の選択肢数が 21 もあると，非常に計算量が大きくなる．ベイジアンネットモデルの構築には，ベイジアンネットワーク構築ソフトウェア Bayonet[本村 03] を用いた．

4.2.3 設問の選択方法

選択する設問数は 5 つとした．設問の選択方法については，以下の 4 種類で実験を行った．

- PC: $ProbChange$ を用いた動的アンケートシステム
- WPC: $W\_ProbChange$ を用いた動的アンケートシステム (与えるエビデンス情報が C の場合のみ使用可能)
- fix_PC: $ProbChange$ を用いた動的アンケートシステムにより，高い頻度で選択される設問 5 つ
- fix_WPC: $W\_ProbChange$ を用いた動的アンケートシステムにより，高い頻度で選択される設問 5 つ (与えるエビデンス情報が C の場合のみ使用可能)

4.3 評価

本実験の評価として，以下の 3 種類の評価指標を用いた．

4.3.1 *Rank-r matching rate*

1 つ目の指標 (以後 *Rank-r matching rate* とする) について説明する．*Rank-r matching rate* は以下の式で表される．

$$Rank\text{-}r\ matching\ rate = \frac{\sum_d \frac{matching(d)}{\min(label(d), recommend)}}{D}$$

ここで，D はユーザー数を，$matching(d)$ はユーザー d において推薦結果と実際の回答が一致した個数を，$label(d)$ はユーザー d の実際の回答結果の個数を，$recommend$ は本実験での推薦個数を表す．実際の回答が推薦数よりも多い場合に，全ての回答をマッチングさせることができないことを考慮した評価指標である．

4.3.2 *Rank-r hit rate*

続いて，2 つ目の指標 (以後 *Rank-r hit rate* とする) について説明する．*Rank-r hit rate* は以下の式で表される．

$$Rank\text{-}r\ hit\ rate = \frac{\sum_d step(matching(d))}{D}$$

$$step(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x = 0) \end{cases}$$

各ユーザーの推薦結果が，実際の回答と一つ以上一致していれば *hit* とし，その率をとった指標である．実際の利用シーンを想定した際に，最も優先度が高く，なおかつ高水準であるべき指標であると考えられる．なお，$recommend = 1$ の場合は，*Rank-r matching rate* と同義である．

4.3.3 *Rank-r nega rate*

3 つ目の指標 (以後 *Rank-r nega rate* とする) について説明する．*Rank-r nega rate* は以下の式で表される．ここでは選択確率が低いものを推薦する．

$$Rank\text{-}r\ nega\ rate = 1 - \frac{\sum_d step(matching(d))}{D}$$

各ユーザーの推薦結果が，実際の回答と一つ以上一致していれば *nega* とし，その率を 1 から引いた指標である．ここでは選択確率が低いものを推薦しているため，実際の回答と一致していなければいないほど良い．

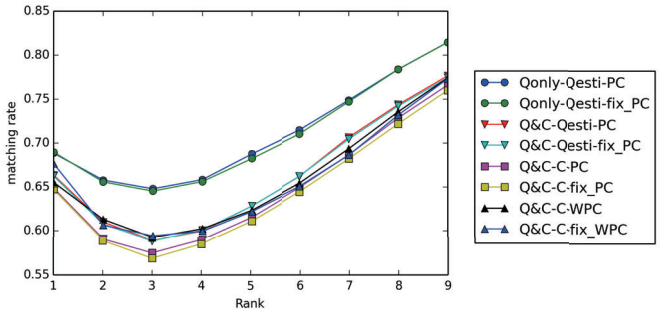

図 3: 各実験条件における *matching rate* の推移

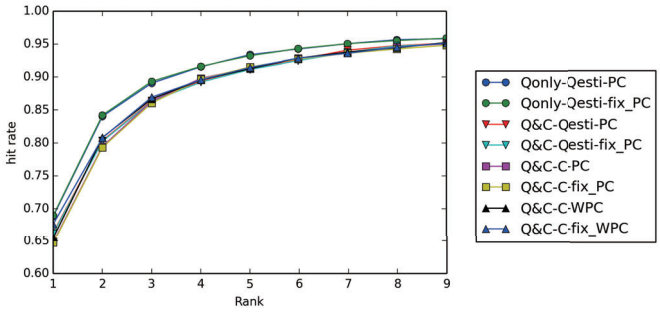

図 4: 各実験条件における *hit rate* の推移

5. 結果・考察

本項では，評価指標を用いて各手法の比較を行った結果，図 3, 図 4, 図 5 のようになった．図 3, 図 4, 図 5 からわかるとおり，基本的に設問のみのベイジアンネットワークにおいて，推定設問情報のエビデンスを与えた場合が評価指標の値が高くなっている．例えば *hit rate* については，セグメント情報を用いずに推定設問情報のエビデンスを与えた手法 "*Qonly-Qesti-PC*" は，提案手法 "*Q&C-C-WPC*" と比較して各ランクで 5 ポイントほど高い精度である．推定設問情報のエビデンスを与えることは，計算量の観点からはそれほど望ましくないが，精度としては最も高いということがわかる．これは，推定設問に効く設問を直接選択してきているため，自然な結果であるといえる．特に "*Qonly-Qesti-fix_PC*" については，事前に設問を確定させるため，設問選定後は動的アンケートとして運用が可能である．しかし，事前の計算が膨大であることが難点である．

また，提案手法である "*Q&C-C-WPC*" については，"*Q&C-C-PC*" と比較した場合は，全体的に高い精度を保っていることがわかる．これは提案手法の効果が出ていると考えられる．とはいえ，推定設問情報のエビデンスを直接与えた場合と比較すると，特に低めの Rank においては 5 ポイントほど精度が落ちている．逆に，セグメントの情報をエビデンスとして与えるという，計算量に配慮した手法の中では，"*Q&C-C-WPC*" 及び "*Q&C-C-fix_WPC*" が，低めの Rank において 2 ポイントほど良い精度を出しているといえる．よって，精度面では推定設問に直接エビデンスを与える手法には及ばないものの，提案手法は総合的に見て有用であると考えられる．

なお，出題する設問を固定するか否かは，それほど評価指標の値に影響していないことが読み取れる．

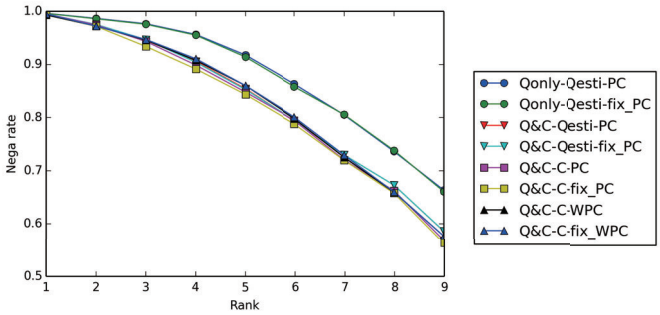

図 5: 各実験条件における *nega rate* の推移

6. まとめ

本研究では，過去に提案した動的アンケートシステムを応用し，ユーザーの設問回答の推定を行うための新しい設問選択アルゴリズムを提案した．具体的には，セグメントから推定設問の確率値変化量を計算し，それを重要度とすることで，間接的に推定設問の変化量が大きくなるような設問を選定するシステムを提案した．そして，設問回答の推定精度比較を行うために，提案手法や，先行研究で使用した設問選択アルゴリズムをそのまま利用した手法，セグメントを用いずに推定設問にエビデンスを直接与える手法を用いて実験を行った．*hit rate* の実験結果より，提案手法はセグメントを用いない手法には 5 ポイント程及ばなかったものの，過去の動的アンケートシステムをそのまま利用した手法と比較するとやや精度が高かった．計算量を考慮に入れると，セグメント情報をエビデンスとして与える方法は非常に有用であり，その方法群の中では提案手法が最も良い精度を達成している．よって，提案手法は非常に有用であるといえる．

今後は，本研究のアルゴリズムを全ての設問に対して適用し，総合的な結果の評価を行うことが必要である．また，推定したい設問に沿ったセグメントを得る方法も求められる．

謝辞

本研究（の一部）は国立研究開発法人科学技術振興機構（JST）の研究成果展開事業「センター・オブ・イノベーション（ＣＯＩ）プログラム」の支援によって行われた．また，本研究は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託事業「人間と相互理解できる次世代人工知能技術の研究開発」の支援を受けて行った．

参考文献

[本村 16] 本村. 第 9 章「ベイジアンネットワークと確率的潜在意味解析による確率的行動モデリング」「確率的グラフィカルモデル」（鈴木譲 他）, 共立出版, 2016

[田村 18] 田村, 櫻井, 本村. Bayesian Network を用いた動的アンケートシステムの提案. 人工知能学会全国大会 1P102, 2018.

[Hofmann 99] T. Hofmann and J. Puzicha, “Latent class models for collaborative filtering”, Proc. 16th international joint conference on Artificial intelligence, 1999

[本村 03] 本村陽一. ベイジアンネットソフトウェア BayoNet. 計測と制御 42.8 (2003): 693-694.