Bayesian Networkを用いた動的アンケートシステムの提案

Adaptive questionnaire system based on Bayesian network

田村 脩^{*1} 櫻井 瑛一^{*2} 本村 陽一^{*3} Shu Tamura Eiichi Sakurai Youichi Motomura

*1*2*3 產業技術総合研究所

National Institute of Advanced Industrial Science and Technology (AIST)

In this study, we aim to find effective and personalized questions from a questionnaire that enable answerers clustering. We propose a question selection method based on a Bayesian network model that is constructed from probabilistic clustering results. We show that the gap between the accuracies of estimating cluster index of our method and the upper bound of them is very small.

1. はじめに

ユーザのモデル化を図り商品のマーケティングや新商品の設 計に生かすことは広く行われてきた. 昨今では、大量のデータの 収集が可能となったことで、客観的なデータからユーザーモデル を作成することが試みられている. その一例としては, 購買履歴 などの人の行動データから行動の類型化を行い、行動の典型的 な例を作成することがある。例えば、「石垣10」では、顧客の購 買行動を Probabilistic Latent Semantic Analysis[Hofmann 99](以 降 PLSA と略す)を拡張したモデルで人の類型化を行い,得ら れた顧客行動のセグメント情報と顧客に対して行ったアンケー トデータに基づいたベイジアンネットワークを作成することで 得られた顧客行動セグメントの構造を明らかにしている.この 論文では、セグメントを数億のログデータから作成している が、その本質は購買商品頻度と顧客という二次元のデータか ら,確率的手法によりセグメントを得ることにある.したがっ て同様の形式のデータであれば、形式的に適用することでセグ メントとその関係を明らかにすることが可能となる. そのよ うな例として,アンケートデータそのものに PLSA にてクラ スタリングを行い,得られたセグメントの特徴をセグメント に対するベイジアンネットワークから見出した研究も存在する [井手 17]. このように分析として、クラスタリング手法とベ イジアンネットワークを組み合わせたセグメント説明モデルに より, データに基づいたクラスタリング結果の解釈が可能とな ることがわかる.

上記の例は分析を目的としていたが,アンケートデータを 拡充するためには,サービス(ゲーミフィケーションなどの形 をふくめ)として,データを蓄積する仕組みが必要である.し かし,次々と未知の人が来るときにその人がどのクラスである かを推定する問題の下で,ベイジアンネットワークを使用して ユーザーモデルを推定する場合,新規の人にアンケートをすべ て回答してもらう必要がある.多くの場合,アンケートは多く の設問から構成され,設問内に多くの選択肢を含むため,すべ てを回答してもらうコストは非常に大きくデータを蓄積する当 初の目論見を損なう恐れが大きい.

そこで、本論文では、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アンケート項目を逐次に提示すること で、ユーザーセグメントの推定をどれだけ可能となるかを検証

連絡先:田村脩,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 2-4-7,0465-35-7815, shu.tamura@aist.go.jp することとした.

提案手法

先に,提案手法で使用する方法を紹介し,提案手法について 述べる.

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 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)を推定をする.

2.2 Bayesian Network

ベイジアンネットとは、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

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

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

と表現できたとすると、その関係はグラフとして図1のように 表現される.このような条件付き確率に基づいたモデルの手段 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となる. そして、アンケート回答結果のデータから、この条件付確率による関係をベイジアンネットワークにより推定し、アンケートの設問とセグメント間の関係性を導き出した.



図 1: ベイジアンネットワークの例

2.3 提案手法

本研究では、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設問の確率値 の変化量をもとに、次の設問を逐次的に選定するシステムを提 案する.

本項では,まず確率値の変化量を算出する際に用いる KL ダイバージェンスについて紹介したのち,逐次的設問選定システムの詳細について述べる.

2.3.1 KL ダイバージェンス

KL ダイバージェンスとは、ある二種類の確率分布がどれだ け似ているかを表す指標である.同じ確率分布では値が0に なるため、便宜的に確率分布間の距離として扱われる. α,βを 離散確率分布とした際の KL ダイバージェンスは、以下の式で あらわされる.

$$D_{KL}(\alpha,\beta) = \sum_{k}^{K} \alpha_k * \log \frac{\alpha_k}{\beta_k}$$
(1)

2.3.2 エビデンス付与による設問の確率値の変化量

セグメントの集合 $C = \{c_1, ..., c_x, ..., c_X\}$ を,設問の集合 $Q = \{Q_1, ..., Q_y, ..., Q_Y\}$ が説明する構造のベイジアンネッ トワークを考える.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 $Q_y = \{q_1, ..., q_k, ..., q_{n_y}\}$ を持つとしたとき, 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 $ProbChange(Q_y)$ を以下の式で定義する.

$$ProbChange(Q_y) = \sum_{q \in Q_y} d(q)$$
 (2)

 $d(q) = \sum_{c \in C} \sum_{x=0}^{1} D_{KL}(new_{-p}(q, c = x), old_{-p}(q))$ (3)

ここで、 $old_p(q)$ は、既に回答された設問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 q の選択確率分布であ る.そして、 $new_p(q,c=x)$ は、既に回答された設問の回答 及び、セグメント c に x(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 q の選択確率分布である.なお、c=0は、クラスタc には属さないという意味であり、c=1は、ク ラスタc に属するという意味である、この2つの分布の KL ダ イバージェンスをとることで、確率値の変化量とした.

2.3.3 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 を持つ選択肢が、優先して選定するべき重要な選択肢であると 考えられる.そこで、既に回答された設問のエビデンスを所与 として、各セグメントにバイナリでエビデンスを与えた際に、 各選択肢の選択確率値の変化の合計値 (*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 preQ とすると, アンケートシス テムで次に質問する設問 nextQ は,以下の式によって決定さ れる.

$$nextQ = Q_{nextqid} \tag{4}$$

$$nextqid = \arg\max_{y|1 \le y \le Y, y \notin preQ} ProbChange(Q_y)$$
(5)

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. 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. 実験

今回の実験では、提案手法である逐次的設問選定を行った場 合と、すべての設問に回答した場合とを比較し、提案手法の有 意性について検討する.

4.1 準備

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,各セグメントを 各設問が説明するようにベイジアンネットワークを作成した. 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 実験概要

本実験の概要について述べる.本実験では,未知のユーザー のセグメント推定を目的としているため,PLSAによって得た セグメントについて,提案手法及び比較手法を用いて推薦結果 を出力し,実際の回答結果と比較した.

また,ユーザーのセグメントに限らず,一人のユーザーが複数の回答を行う設問についても推論が可能であるため,重要な 情報であると考えられる「車の購入に対する重視点」を示す設 問についても予測を行うこととした.

データセットとしては、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.

まず,提案手法については,回答する設問数を5つとした. 評価の際には,動的に採用した5つの設問をエビデンスとして,ベイジアンネットワークで予測を行い,確率の高い上位1 件または3件を推薦結果として出力した.

また,比較手法として,以下の二種類の実験も行った.

・全 30 問の設問の回答結果をエビデンスとし、ベイジアンネットワークで予測

・全 30 問の設問の回答結果及び, PLSA によって得られたセグ メントをエビデンスとし、ベイジアンネットワークで予測(車 の購入に対する重視点の予測時のみ)

4.3 評価

本実験の評価として,以下の2種類の評価指標を用いた. ・matching_rate:ユーザー単位での実際の回答と推薦結果の 一致率を算出し,全ユーザーについて平均をとった指標.

・*hit_rate*: 推薦結果と実際の回答で一致するものが1つ以上 存在すれば成功であるという考えによる指標. 以下に詳細を説明する.

4.3.1 matching_rate

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

$$matching_rate = \frac{\sum_{d} \frac{matching(d)}{\min(label(d), recommend)}}{D}$$
(6)

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 (1 または 3) 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 hit_rate

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

$$hit_rate = \frac{\sum_{d} step(matching(d))}{D}$$
(7)

$$step(x) = 1: x > 0 \tag{8}$$

$$0:x \leq 0$$

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、matching_rate と同義であるため、recommend = 3 の場合について評価する.また、セグメントの予測において は、各ユーザーは一つのセグメントにのみ所属しているので、 recommend = 3 の場合についても matching_rate と同義で あるため、hit_rate の評価は行わない.

結果・考察

本項では,まず評価指標を用いて各手法の比較を行い,次 に動的設問選択によってどのように設問が選ばれたのかを把握 する.

5.1 各手法の比較

5.1.1 PLSA を用いたユーザーセグメントについての評価

PLSA を用いたユーザーセグメントについて, 各手法の比較 を行う.

表1に,各手法における評価指標の一覧を示す.なお,表1 の評価指標の前に表示されている"1"や"3"は,各ユーザーに対 する推薦個数である recommend の値を表している.

表 1: セグメントについての評価比較

	1_matching_rate 3_matching_rate	
提案手法	51.11%	87.99%
全設問	68.78%	94.49%



図 2: 採用された設問の採用頻度

表1より,推薦数が1つの場合における matching_rate に ついては,提案手法と最高評価の手法の間で,約18 ポイント の差にとどまっていることがわかる.同様に,推薦数が3つの 場合の matching_rate についても約7 ポイントの差となって いる.

5.1.2 車の購入に対する重視点についての評価 車の購入に対する重視点について、各手法の比較を行う。

表 2 に,各手法における評価指標の一覧を示す.

表 2: 車の購入に対する重視点についての評価比較

	1_matching_rate	3_matching_rate	3_hit_rate
提案手法	64.57%	57.94%	85.98%
全設問	71.75%	67.16%	90.78%
全設問及び			
セグメント	72.49%	67.46%	90.84%

表2より,推薦数が1つの場合における matching_rate に ついては,提案手法と最高評価の手法の間で,約8ポイント の差にとどまっていることがわかる. 同様に,推薦数が3つ の場合の matching_rate については約10ポイント, hit_rate については約5ポイントの差となっている.

5.2 動的設問選択による設問リスト

動的な設問選択によって,扱われる設問の幅がどれ程広がったか,確認を行う.図2に,提案手法におけるセグメント予測について,採用された設問と,その採用頻度について示す.

図2より、5つの設問に答える提案手法であるにも関わらず、 16もの幅広い設問から回答を得ていることが読み取れる.これは、前に回答された設問によって、次の設問が変化している ことの裏付けであるといえる.

6. まとめ

今回の実験により,動的設問選択を用いた少数設問による 予測は,全設問の回答を用いた場合と比較してもそれほど劣っ ていないといえる.今後は,設問選択における選択基準につい て,より深く考察をすべきである.

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(C OI)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

- [石垣 10] 石垣, 竹中, 本村. 日常購買行動に関する大規模デー タを融合したベイジアンネットユーザモデル. 人工知能学 会全国大会 3J1-NFC1a-2, 2010.
- [井手 17] 井手,川本,山下,本村. ミクロアグリゲーションを用 いた匿名化による確率的潜在空間意味解析. 人工知能学会 全国大会 1K2-2, 2017.
- [Hofmann 99] T. Hofmann and J. Puzicha, "Latent class models for collaborative filtering", Proc. 16th international joint conference on Artificial intelligence, 1999
- [櫻井 16] 櫻井, 本村, 安松, 坂本, 道田. 感性ユーザーモデルの 構築のためのデータ収集方法. 日本行動計量学会第 44 回 大会予稿集, 2016
- [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.