Web 閲覧履歴に基づく消費者セグメント特定のための属性ラベル学習モデル

経営情報学研究　　5219F001-3　青木章悟
指導教員　後藤正幸

Web Browsing History-based Systematic Method of Attribute Labeling for Consumer Segment Targeting

AOKI Shogo

1. はじめに

多様な製品が溢れる現在においては，自社の製品がターゲットとすべき消費者セグメントを明確にすることが重要である．全消費者のうちからターゲットとする消費者セグメントを特定するために，例えば「30 代の男性会社員」などのようにターゲット属性を用いて特定することは常套手段の一つである．しかし，全消費者の消費者属性と趣味嗜好を含むデータがあれば分析的な観点からターゲット属性を特定できるが，そのようなデータを各企業が独自に収集することは膨大なコストが掛かり現実的ではない．そのため，多種多様なサンプル消費者の Web 閲覧履歴や購買履歴を保有するコンサルティング企業に依頼する形で，サンプル消費者のデータ分析を通じてターゲットを特定することが一般的に行われている．特に，サンプル消費者の Web 閲覧履歴などの行動履歴データをクラスタリングし，ターゲットとして相応しいクラスタを特徴付ける属性（クラスタ属性）を発見する方法はよく用いられるアプローチの一つである．

しかし，クラスタ属性はしばしばクラスタに所属するサンプル消費者の属性統計量などから分析者の判断によって付与されることが多い．その上，一つのクラスタに複数のクラスタ属性を想定しシナリオ別の施策が検討されるようなこともある．このような定性的な分析や判断が入る場合，クラスタ属性の選定は分析者の経験やスキルに強く依存してしまう．そのため，消費者嗜好が反映された過去のサンプル行動データのクラスタリング結果から，客観的な評価基準に基づいてクラスタ属性を特定可能な数理モデルがあれば，分析者の作業や施策検討を強力に支援できると考えられる．

そこで本研究では，多様な趣味嗜好を含む Web 閲覧履歴からサンプル消費者をクラスタリングし，各クラスタに対して，ターゲットとして有効なクラスタ属性を付与するための最適化モデルを定式化する．さらに，与えられた目的関数に対して最適なターゲット属性を探索する手法を提案する．最後に，実データを用いて提案手法の有効性を示す．

2. 問題設定

本研究で対象とする問題は，取得された Web 閲覧履歴によってサンプル消費者をクラスタリングし，各クラスタに適切なクラスタ属性を付与することで，Web 閲覧履歴が取得できていない全消費者に対しても同様の嗜好を持つであろうセグメントを推定するという問題設定である．ここでクラスタ属性とは,「性別」や「年齢」などの属性変数を AND 記号 ($\wedge$) を用いて「30 代 $\wedge$ 男性 $\wedge$ 会社員」のように組み合わせることでクラスタの集合を適切に表したものと定義する．この分析を行うにあたり，使用するデータと従来行われてきたターゲット属性特定のための分析プロセスを以下に示す．

2.1. 使用データ

本研究で想定する問題設定では，多様なカテゴリの Web サイトに対するサンプル消費者の Web 閲覧履歴を使用する．各サンプル消費者 $n = 1, \cdots, N$ には閲覧した Web サイトの URL が時系列で与えられており，さらに「性別」や「年齢」などの属性変数も与えられるものとする．

2.2. ベースとなる分析プロセス

本研究で対象とする分析プロセスの手順を述べる．従来，Web 閲覧履歴を用いたマーケティング分析では，サンプル消費者を適当なモデル $\mathcal{C}$ でクラスタリングし，各クラスタを所属データの消費者属性で特徴付ける方法は基本的アプローチである．この分析プロセスの手順を以下に示す．

手順 1 Web 閲覧履歴に対し，適当なクラスタリングモデル $\mathcal{C}$ を用いてサンプル消費者をクラスタリングする．

手順 2 各クラスタ $k = 1, \cdots, K$ に対し，所属するサンプル消費者を抽出し消費者グループ $\boldsymbol{U}_k$ とする．確率的なソフトクラスタリングの場合は，所属確率が ε 以上のサンプル消費者を $\boldsymbol{U}_k$ とする．

手順 3 消費者グループ $\boldsymbol{U}_k$ から何らかの方法でクラスタを特徴付ける要素属性 $x_1^k x_2^k \cdots x_{D_k}^k$ を発見し，これらを AND($\wedge$) で結びつけて，クラスタ属性 $\boldsymbol{X}_k = x_1^k \wedge x_2^k \wedge \cdots \wedge x_{D_k}^k$ としてクラスタ k に付与する．

上記の手順により，分析者はクラスタを特徴付けるクラスタ属性を選択し，依頼者に分析結果を提供する．

3. 提案手法

2.2 節で示した分析プロセスでは，前述の手順 3 において消費者属性の統計量等を用いて分析者が独自にクラスタを特徴付けるクラスタ属性を付与していたが，これには明確な基準がなく，クラスタ属性の決定には分析者の経験やスキルに強く依存してしまう．そこで本章では，手順 3 の改善として，クラスタ属性決定の際の曖昧な評価基準を明確化し，評価指標が高くなるようにクラスタ k のクラスタ属性 $\boldsymbol{X}_k$ を選択するための手法を提案する．

3.1. 評価指標

クラスタ k に対して付与されたクラスタ属性 $\boldsymbol{X}_k$ の適切性を議論するため，精度と網羅性の二つの指標を考える．ここで精度とは，付与したクラスタ属性を持つ全サンプル消費者のうち，実際にクラスタ k へ所属するサンプル消費者の割

合とする．また，網羅性は，クラスタ k に所属する全サンプル消費者のうち，当該クラスタ属性が付与されたサンプル消費者の割合とする．精度のみが高くなるよう属性を付与すると，付与するクラスタ属性を複雑にすることで精度を高められるため，各クラスタに付与される属性ラベルが複雑になり，特徴の把握は困難となる．他方，網羅性のみが高まるようクラスタ属性を付与する場合，全ての属性変数を選択すれば良いことになってしまう．そこで，精度と網羅性のトレードオフを考慮し，これらの調和平均である F 値を評価基準とする．そして，この指標が高い属性を最適な属性と定義する．

3.2. 属性選択手法

3.2.1. 単一のクラスタ属性選択 (提案手法 1)

手順 3 で最適なクラスタ属性を推定するために全ての属性変数の組み合わせを全数探索すると，計算量は属性変数の数の指数オーダーとなり，属性変数が多い場合，現実的ではない．そこで，遺伝的アルゴリズム (以下，GA)[1] を用いて要素属性の選択を最適化問題として解くことを考える．

GA を援用してクラスタを最もよく表すクラスタ属性を推定するためには，F 値が最も高くなるクラスタ属性を選択すればよい．従って，GA の個体をクラスタ属性，目的関数を F 値とすることで，クラスタ内の F 値が高くなるようクラスタ属性を最適化する．提案手法における個体のイメージを図 1 に示す．

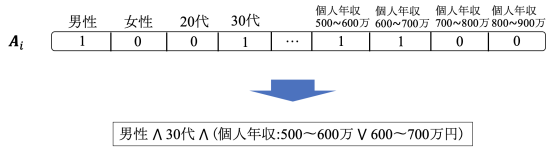

図 1: 消費者属性を適用した個体のイメージ

ここで，系列長 L の個体 i を $\boldsymbol{A}_i = (a_1^i, a_2^i, \cdots, a_L^i)$ と定義したとき，$a_l^i \in \{0, 1\}$ は l 番目の属性を表す．すなわち，$a_l^i = 1$ で要素属性として属性 l を選択し，$a_l^i = 0$ で属性 l を選択しないことを表現する．また，図 1 の「個人年収」という属性変数に示すように，個体 $\boldsymbol{A}_i$ は，属性変数間で AND($\wedge$) を取る一方，属性変数内では属性値を OR($\vee$) を用いて連結できるものとし，クラスタ属性を表現する．そして，個体数が G である集団 $\mathcal{A} = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_G\}$ から F 値が最も高くなる個体を選択し，クラスタ属性として推定する．ただし，GA は各クラスタについて独立に適用する．

3.2.2. 二つのクラスタ属性選択 (提案手法 2)

提案手法 1 を Web 閲覧履歴に適用することで，最適なクラスタ属性を推定することが可能になると考えられる．しかし，一つのクラスタに複数の特徴的なクラスタ属性が存在する場合，単一のクラスタ属性 $\boldsymbol{X}_k$ のみではクラスタを適切に表現できているとは言い難い．そこで提案手法 2 では，クラスタ内に二つの特徴的なクラスタ属性があることを仮定し，二つのクラスタ属性を推定する方法を提案する．

クラスタを特徴付ける二つ目のクラスタ属性 $\tilde{\boldsymbol{X}}_k$ には，評価値が高いうえに一つ目のクラスタ属性との差異が大きいものが選択されることが望ましい．ここで GA の問題点として，最適解を探索するうちに，集団中の遺伝子が偏った局所解に収束してしまうフォーカシング問題が挙げられる．そのため，個体の評価値が高くクラスタ属性間の差異が大きくなる組み合わせが複数存在する多様性のある集団を維持することは困難である．そこで提案手法 2 では，評価値が高く多様性のある集団を維持するニッチング法を適用することを考える．

ニッチング法の代表的な手法として Fitness Sharing [2] が知られている．Fitness Sharing は，集団における自身の特異性を評価値に重み付けすることで多様性を得ることを試みる．GA の個体 i の評価値を f_i とし，重み付け後の評価値を f_i' とすると，Fitness Sharing による重み付け評価値は以下のように定義される．

$$f_i' = \frac{f_i}{m_i} = \frac{f_i}{\sum_{j=1}^{G} sh(d_{ij})} \tag{1}$$

ただし，m_i は個体 i の集団中における特異度であり，自身と評価値を共有する個体のおおよその数を表している．また，d_{ij} は個体 i と j の距離を表す．関数 sh はシェアリング関数と呼ばれ，式 (2) で定義される．

$$sh(d_{ij}) = \begin{cases} 1 - \left(\frac{d_{ij}}{\sigma}\right)^{\alpha}, & d < \sigma \\ 0, & otherwise \end{cases} \tag{2}$$

ただし，σ は個体間の非類似性を示す閾値であり，α は距離が及ぼす影響の度合いを調整するパラメータである．GA の評価値 f_i を f_i' に変換することで，多様性を保つよう集団を更新する．また，本手法における個体間の距離 d_{ij} は，二つのクラスタ属性のどちらかに所属するサンプル消費者のうち，どちらにも所属するサンプル消費者の割合に従うものと定義する．すなわち，GA の個体 i に該当するサンプル消費者集合を $\mathcal{S}^{(i)}$ として，以下のように計算する．

$$d_{ij} = 1 - \frac{|\mathcal{S}^{(i)} \cap \mathcal{S}^{(j)}|}{|\mathcal{S}^{(i)} \cup \mathcal{S}^{(j)}|} \tag{3}$$

多様性のある集団を形成した後，集団における全通りの個体のペアについて F 値を計算することで，最も高い F 値となる二つの個体を最終的なクラスタ属性として選択する．これらに対する F 値は，当該クラスタ属性を持つサンプル消費者として $\boldsymbol{X}_k$ または $\tilde{\boldsymbol{X}}_k(\boldsymbol{X}_k \cup \tilde{\boldsymbol{X}}_k)$ に当てはまるサンプル消費者を用いることで計算する．

4. 実データ分析

4.1. クラスタリング手法の選定

データをソフトクラスタリングする手法として，トピックモデルの一種である Latent Dirichlet Allocation(以下，LDA) [3] が知られている．LDA は本来，文書の生成過程を表現する生成モデルとして提案されたマルチトピックモデルである．学習データに存在しない新規文書に対する各クラスタへの所属確率が予測可能であり，ベイズモデルをベースにしているため過学習を起こしにくいという特性を持つ．本研究においても，サンプル消費者をクラスタリングするモデル $\mathcal{C}$ として LDA を採用する．

表 1: 提案手法 1 で推定したクラスタ属性に対して F 値が高かった上位二つのトピック

k	精度	網羅性	F 値	クラスタ属性	共起サイト (上位 6 件)
2	0.230	0.733	0.350	女性 ∧(アルバイト∨ 会社員 (一般社員)∨会社員 (管理職)∨ 会社経営 ∨公務員 ∨ 専業主婦 ∨ 専門職∨ 契約社員 ∨自営業 ∨ その他)∧ (個人年収:〜900 万 ∨不明)	ユニクロ, @cosme, Qoo10(インターネット通販サイト), LOHACO, ベルメゾンネット, ホットペッパービューティー
14	0.289	0.336	0.311	男性 ∧20〜30 代∧ (アルバイト ∨会社員 (一般社員)∨ 公務員 ∨学生 ∨ 契約社員 ∨ 自営業∨ その他)∧未婚 ∧ 子無し	ニコニコ動画, ニコニコ静画, ニコニコ生放送, アダルトサイト, 同人作品ショップ Dlsite, 無料レンタル Wiki サービス

4.2. 未学習の消費者に対するクラスタ属性の整合性評価

選択したクラスタ属性を実際にターゲットとする際，閲覧履歴を保有しない全消費者に対しての適切性も評価をする必要がある．そこで，サンプル消費者を学習データとテストデータに分割し，学習データから推定したクラスタ属性を未観測の消費者に見立てたテストデータに対し評価する．具体的には，学習データにより学習した LDA のパラメータからテストデータの各クラスタに対する所属確率を算出し，学習データから GA で付与したクラスタ属性が所属確率が ε 以上となるテストデータ中の対象消費者グループをどの程度正しく推定できるかを評価する．

4.3. 分析条件

提案手法の有効性を確認するため，株式会社ヴァリューズ提供の Web 閲覧履歴を用いてモデルの評価及び分析を行う．データ期間は 2019 年 2 月 1 日から 2019 年 4 月 30 日であり，総閲覧数は 138,765,006 件，サイト数は $M = 7,351$ である．サンプル消費者は，$N_{train} = 34,308$ 名の学習データ，$N_{test} = 8,577$ 名のテストデータにランダムに分割して実験に用いる．サイトは URL のドメイン名を使用する．属性変数は「性別」,「年齢」,「職業」,「子供の有無」,「未既婚」,「世帯年収」,「個人年収」を使用し，全て離散データとする．ただし「年齢」は 20 代〜70 代と 80 代以上の 7 段階とし，「職業」は「その他」を含めて 14 段階,「年収」は「不明」を含め 11 段階に離散化した．LDA のクラスタ (トピック) 数は $K = 30$ とし，学習には変分ベイズ法を使用し，学習回数は 200，ハイパーパラメータの初期値は全て 0.1 とした．GA の個体数は 300 とし，学習回数は 500，選択方式はルーレット選択，交叉方式は交叉確率 0.8 の二点交叉とし，突然変異方式は突然変異確率 0.1 の置換方式とする．提案手法 2 におけるパラメータについては，$\sigma = 0.01$，$\alpha = 3$ とする．また，属性付与のための閾値は $\varepsilon = 0.2$ とした．

4.4. 分析結果

4.4.1. 提案手法 1 の分析結果と考察

提案手法 1 により推定されたクラスタ属性の各トピックに対する F 値を図 2 に，そのうち F 値が最も高い上位二つのトピックの詳細を表 1 に示す．ここで，表 1 のクラスタ属性は，表中に記載された「性別」や「年齢」など全ての属性変数について，いずれかに当てはまるサンプル消費者を指す．

図 2 より，トピック間で評価指標に大きく差異があることが確認できる．これは，データ中の属性変数だけでトピックを表現する難易度がトピックに応じて異なるためであると考えられる．ここで，評価指標が全体で 2 番目に高いトピック 14 について考察する．全学習データのサンプル消費者のうち，トピック 14 への所属確率が $\varepsilon(= 0.2)$ 以上であるサンプル消費者の分布を確認するため，各トピックへの所属確率分布によりサンプル消費者を 30 次元特徴ベクトルで表し，t-SNE [4] を適用して 2 次元に圧縮したうえで該当するサンプル消費者を濃い色でプロットした．トピック 14 に所属するサンプル消費者の分布と提案手法 1 により推定されたクラスタ属性に当てはまるサンプル消費者を図 3 に示す．図 3 から，推定した属性のサンプル消費者がトピック 14 に所属するサンプル消費者を中心に分布していることがわかる．また表 1 の $k = 14$ に示す学習データに対する精度，網羅性は共に F 値に近い値であることから，精度と網羅性のどちらか一方が高くなることでもう一方が低くなることを防止でき，提案手法により適切なクラスタ属性が推定できていると言える．

他方，最も F 値が高いトピック 2 のクラスタ属性を推定した結果，精度と比較して網羅性が非常に高くなった．このトピックに共起するサイトは表 1 の通りであり，推定されたクラスタ属性から家族の形態に問わず多くの女性がこの

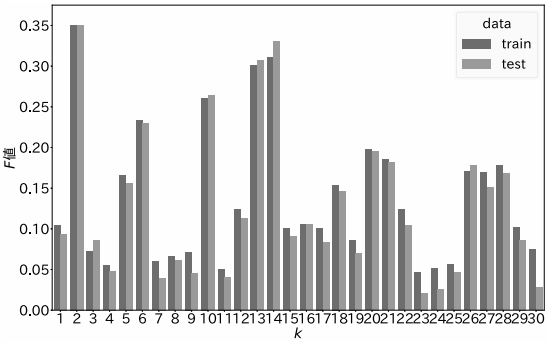

図 2: 推定したクラスタ属性の各トピックに対する評価指標

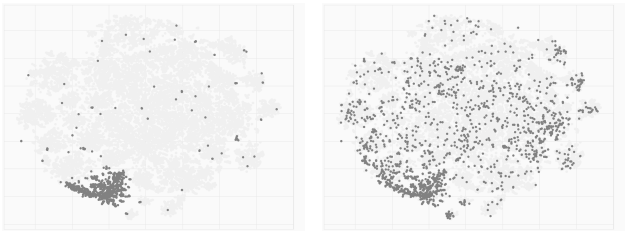
(a) トピック 14 に確率 ε 以上所属するサンプル消費者　(b) 提案手法 1 で推定したクラスタ属性に該当するサンプル消費者

図 3: t-SNE を用いたサンプル消費者の分布

表 2: 提案手法 2 で推定されたトピック 2 と 14 のクラスタ属性

k	精度	網羅性	F 値	属性 1	属性 2
2	0.238	0.645	0.347	女性 ∧30〜60 代 ∧ (アルバイト∨ 会社員 (一般社員)∨会社経営 ∨ 公務員 ∨専業主婦 ∨ 専門職 ∨ 無職 ∨ 自営業 ∨その他) ∧ (世帯年収 : 〜300 万 400〜1,500 万)∧(個人年収 : 〜200 万∨ 300 万台 ∨600〜800 万 ∨ 不明 }	女性 ∧30〜60 代 ∧ (アルバイト∨ 会社員 (一般社員)∨公務員 ∨ 専業主婦∨ 契約社員 ∨無職 ∨ 自営業 ∨ その他) ∧(世帯年収 : 〜700 万 ∨ 1,000 万〜 ∨不明) ∧ (個人年収 : 〜300 万∨ 400〜1,000 万 ∨不明)
14	0.224	0.431	0.295	男性 ∧20〜40 代 ∧ (アルバイト∨ フリーランス ∨会社員 (一般社員)∨ 会社経営 ∨公務員 ∨ 学生 ∨ 専業主婦∨ 無職 ∨自営業 ∨ その他) ∧ 未婚∧ (世帯年収:〜500 万 ∨700〜900 万∨ 1,500 万〜∨不明) ∧ (個人年収 : 〜600 万∨ 700 万〜∨不明)	男性 ∧20〜40 代 ∧ (アルバイト∨ フリーランス ∨会社員 (一般社員) ∨ 学生∨ 専業主婦 ∨その他) ∧ 未婚∧ (世帯年収:〜200 万 ∨300〜900 万∨ 1,000〜1,500 万 ∨不明)∧ (個人年収:〜200 万 ∨300〜600 万∨ 1,000〜1,500 万 ∨不明)

トピックに興味を示していることがわかる．提案手法で推定したクラスタ属性に対してマーケティングを行うことで，このトピックに興味を持つ多くの消費者に対してアクセスできるが，精度を考慮すると，興味を持たない消費者に対してもマーケティングを行うことになるため，膨大な予算をかけた場合に費用対効果が低くなる可能性が示唆される．

4.4.2. 提案手法 2 の分析結果と考察

続いて，提案手法 2 により各トピックについて二つのクラスタ属性を推定した結果を示す．まず，推定されたクラスタ属性の各トピックに対する F 値を図 4 に示す．

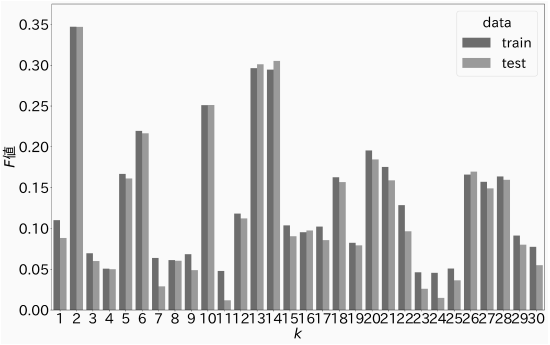

図 4: 推定した 2 属性の各トピックに対する評価指標

図 2 と図 4 を比較すると，わずかに提案手法 2 の F 値が低下することが確認された．ここで，表 2 に示すトピック 2 で推定された二つのクラスタ属性を見ると，提案手法 1 で推定されたもの（表 1）を細分化するように推定された．例えば職業については，クラスタ属性 1，2 で共通する項目は「アルバイト」,「会社員 (一般社員)」,「公務員」,「専業主婦」,「無職」,「自営業」であり，どちらかにのみ存在する項目は「会社経営」,「専門職」と「契約社員」であった．つまり，本トピックに興味を持つ主な消費者の職業は 2 属性で共通する職業であり，このグループを二つに分割するとしたら,「会社経営」,「専門職」である消費者と「契約社員」である消費者に分けられるという結果になったと考えられる．これらの消費者についてより詳細に分析をすることでターゲット属性を決定するための一助となる可能性がある．

5. 考察

分析結果から，提案手法 1，2 により各トピックに興味がある消費者へのターゲティングが有効であると考えられるターゲット属性を特定できることが期待される．しかし，トピック 2 のように，精度より網羅性が著しく高いトピックが多く確認された．これは，GA で F 値を最大化する際に，精度より網羅性を高める方が容易であるためと考えられる．このようなトピックに所属する消費者をターゲティングする場合，より多くの消費者に対してアプローチしたい場合は推定されたクラスタ属性をターゲットとすることは有効であると考えられるが，予算が小さく消費者をより正確にターゲティングしたい場合にこれらの属性を用いてアプローチすることは望ましくない．そのような場合，網羅性と精度の重要性のバランスを考慮できる指標である F_β 値を目的関数に用いることで，目的に見合った属性を推定できる可能性がある．

6. まとめと今後の課題

本研究では，マーケティング対象とすべきターゲット属性をサンプル消費者から付与するため，閲覧傾向が類似したクラスタ内から適切なクラスタ属性を GA により推定しターゲットとするための方法論と属性の適切性を評価する指標を提案した．また，提案手法により推定したクラスタ属性が全消費者に対して適合するか否かを評価できる可能性を示した．これらにより，分析者の経験やスキルに依存する作業や施策検討を支援できることが期待される．

本研究では，デモグラフィック属性のみを用いてターゲット属性を付与する手法を提案した．しかし実際は，この他にサイコグラフィック属性なども用いて総合的に判断することで最終的なターゲット属性を決定することが多いため，これを考慮した手法の検討については今後の課題とする．

参考文献

[1] Y. Kim, N. Street, G. Russell, and F. Menczer, “Customer targeting: A neural network approach guided by genetic algorithms,” *Management Science*, vol. 51, pp. 264–276, 02 2005.

[2] B. Sareni and L. Krahenbuhl, “Fitness sharing and niching methods revisited,” *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 3, pp. 97–106, 1998.

[3] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 05 2003.

[4] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.