

レビューサイトのプロフィール推定とそのマーケティングへの応用

永田 弘祐[†] 永田 靖[‡]

[†]早稲田大学大学院 創造理工学研究科 経営システム工学専攻 [‡]早稲田大学 創造理工学部 経営システム工学科

1. 研究の背景と目的

近年、ユーザは製品に対する評価をレビューとして書き込み、企業側はレビュー情報をマーケティングに活用している。ただ、一部のユーザは、自身の性別や年齢等のプロフィール情報をレビュー上で公開しない。本研究では、このようなプロフィール情報が欠損しているデータと欠損していないデータをそれぞれ不完全データ、完全データと呼ぶ。また、完全データと不完全データを合わせたデータを統合データと呼ぶ。企業のマーケティングが性別等のセグメントごとに、製品の評価を得ることを目的にしている場合、不完全データは解析に活用できない。

そこで、本研究では不完全データのプロフィールを推定する方法を提示する。同時に、欠損しているプロフィールを推定した後に、完全データと統合データから得られる知見が異なるか検討することを研究目的とする。

2. 完全データと不完全データの評価値比較

楽天市場[1]における異なる 16 種類の製品のレビューに記載された製品の総合評価の平均値を完全データと不完全データに分けて集計した。その結果を表 1 に示す。表 1 において、不完全データの評価値は、完全データの評価値と比較して相対的に低いことがわかる。

表 1. 完全データと不完全データの評価値比較

製品情報		総合評価		サンプル数	
大カテゴリ	小カテゴリ	完全	不完全	完全	不完全
衣類	レディースファッション	4.436	4.274	567	848
	メンズファッション	3.955	3.892	401	415
	靴	4.332	4.233	383	1,028
食品	バッグ	4.261	4.218	353	1,081
	食品	4.415	4.145	804	490
	和菓子	4.649	4.608	844	576
	ソフトドリンク	4.792	4.741	520	850
	ビール	4.677	4.638	895	563
電子機器	ノートPC	4.505	4.407	305	427
	DVDプレーヤー	4.208	4.202	207	420
	掃除機	3.670	3.530	699	932
	電動歯ブラシ	4.100	4.070	1,553	2,354
日用品	布団	4.562	4.488	559	882
	寝具	4.463	4.410	562	773
	食器	4.550	4.504	605	671
	洗剤	4.653	4.663	804	629

3. 従来研究

3.1. プロフィール推定

テキストデータのプロフィール推定は、ブログなどを対象に行われてきたものの、推定結果をマーケティング

Profiling Estimation for Customer Review and its Application for the Marketing

[†]Kosuke Nagata [‡] Yasushi Nagata

[†]Department of Industrial and Management Systems Engineering, Graduate School of Creative Science and Engineering, Waseda University

[‡]Department of Industrial and Management Systems Engineering, School of Creative Science and Engineering, Waseda University

にどのように応用するか具体的に示した研究は少ない。

これに対して、池田ら[2]は、Twitter を対象にプロフィールの推定を行い、推定結果を新製品の企画などへ応用できることを示している。なお、池田ら[2]は、投稿者の過去の投稿内容を遡って取得しているものの、本研究で扱った楽天市場[1]の不完全データでは投稿者の過去の投稿内容を得ることはできない。つまり、本研究では単一のレビューからプロフィールを推定することが求められるため、Twitter やブログなどと比較すると、テキスト自体から得られる情報が少ないという問題がある。

3.2. 欠損のメカニズム

いま、2 変量データ $(x_1, x_2)^T$ において、欠損が x_2 に生じるとき、欠損が x_2 の値および x_1 の値に依存するとき、欠損は無視できない[3]。本研究に適用すると、2 章の結果から、評価値 x_1 が低い場合にプロフィール x_2 を公開しない可能性がある。また、欠損はプロフィール情報 x_2 に依存する可能性もある（男性の方がプロフィールを公開しない等）。このことから、本研究で扱うプロフィール情報の欠損は無視できないと考えられる。

4. 研究内容

4.1. 用いたデータの概要

楽天市場[1]より、4 種類のサイクロン掃除機のデータを取得した。サンプル数の内訳を表 2 に示す。これら 4 製品においても、不完全データの評価値は、完全データと比較して相対的に低く、欠損は無視できないことがわかった。なお、本研究で推定の対象としたプロフィール情報は、性別（男女の 2 値）である。

表 2. 各掃除機のサンプル数内訳

製品	完全		不完全
	男性	女性	
製品A	548	215	410
製品B	305	280	750
製品C	311	225	653
製品D	144	175	372

4.2. レビュー内に含まれる形態素の件数を集計

レビュー内に含まれる文書を形態素解析によって単語分割し、各レビュー内に形態素が含まれていれば 1、含まれていなければ 0 として、0-1 行列を集計する。

4.3. 推定に用いる変数の選択

男女間で用いられる特徴的な形態素を抽出するために、フィッシャーの正確確率検定を用いる。検定の結果を基に、P 値の小さい上位 50 語を後に行う推定で利用する。P 値の小さい上位 5 語を例として、表 3 に示す。表 3 において、*印が付いている形態素は、4 製品中 2 製品以上の掃除機で P 値の小さい上位 50 語に選ばれた形態素

である。つまり、*印が付いていない形態素は、製品固有の男女のレビューで差が出る形態素、*印が付いている形態素は製品共通の男女のレビューで差が出る形態素と言える。本研究では、複数の製品をまとめてではなく、個々の製品ごとに推定に用いる識別器を作成する。これはここで示したように、製品固有の形態素と製品共通の形態素を抽出するためである。製品固有の形態素と製品共通の形態素の両方を用いて性別を推定する方が製品共通の形態素のみを用いて推定するよりも、推定精度が高くなると見込まれる。

表3. 掃除機4製品におけるP値の小さい上位5語

製品A		製品B		製品C		製品D	
形態素	P値	形態素	P値	形態素	P値	形態素	P値
です	0.000	重い*	0.000	私*	0.000	かける*	0.000
私*	0.000	ホース	0.000	まし*	0.001	私*	0.000
わい*	0.000	!*	0.000	母*	0.001	妻*	0.000
思う*	0.000	まし*	0.001	届く	0.002	デザイン	0.000
入*	0.000	可愛い*	0.001	買える	0.002	毎日*	0.001

これとは別に、本研究ではレビュー内に含まれる製品の使い道・製品の使用者・製品の購入頻度、レビューの文字数の情報を推定のための説明変数として用いる。

4.4. 完全データを用いた性別の推定

推定手法として、Support Vector Machine(SVM)を使用した。用いたカーネルは線形カーネルで、適宜、コストマージンパラメータ C を設定した。

なお、SVM の分離超平面から近いサンプルは、4.3 節で示した特徴的な形態素がレビュー文書中に少ないなどの理由で、曖昧な判別が行われたサンプルであり、判別の信頼度が低いデータと言える。そこで、分離超平面からの距離に閾値を設け、閾値が一定の値以下のデータを判別不能と判断するとき、どの程度の精度で判別できるか検証する。4 製品それぞれで閾値を変化させたときの適合率と再現率の推移を表4に示す。表4において、閾値を変化させない場合は、製品A以外の3製品において、適合率、再現率はそれぞれ約70%程度で推定できている。また、閾値が0.5のときに、製品A以外の3製品において適合率80%程度、再現率60%程度で推定できていることがわかる。製品Aについては、他の3製品と比較して、男女間のサンプル数が不均衡であったために、男女間で両指標に差が出たと考えられる。

4.5. 欠損値補完前後での評価値の比較

性別の情報が欠損していない完全データを基に作成した識別器を用いて、不完全データの性別を推定し、欠損値補完前後での評価値の比較を行う。この際、完全データであれば、男女それぞれのサンプル数を1と見なすことができるが、不完全データでは推定精度の低いサン

プルも存在するので、単純に1と見なすことは妥当でないと考えられる。そこで、不完全データであるサンプル i が男性に属する所属確率を $P(y_i = \text{男性} | x_i)$ 、女性に属する所属確率を $P(y_i = \text{女性} | x_i)$ とし、所属確率をサンプル数と見なす。

それぞれのサンプルの各群に対する所属確率を求め、欠損値補完前後での完全データと統合データの評価値を比較した。例として、サイクロン掃除機Dの比較結果を表5に示す。表5において*印が付いているのは、完全データと統合データそれぞれの男女間において評価値の低い方の群である。表5における掃除機Dの総合評価・サイズ・静音性の項目に加えて、掃除機Aの総合評価、掃除機Cのデザインの項目において、完全データと統合データの間で評価値の低い群が変化した。これは、メーカーが完全データのみで評価値が低い群を判断した場合と統合データで判断した場合とで結果が全く逆になる場合があることを示している。

表5. 欠損値補完前後での評価値の比較結果(掃除機D)

評価項目	評価値				
	完全データ		統合データ		不完全データ
	男性	女性	男性	女性	
総合評価	4.653	4.606*	4.561*	4.578	4.522
デザイン	4.632*	4.651	4.579*	4.611	4.556
手入れ	4.417	4.309*	4.316	4.267*	4.231
サイズ	4.563	4.549*	4.503*	4.504	4.460
静音性	3.597	3.554*	3.476*	3.516	3.433
パワー	4.604*	4.617	4.571*	4.612	4.578
使いやすさ	4.458*	4.537	4.410*	4.461	4.382

5. 結論と今後の課題

本研究では、製品のレビューに対するプロフィール推定方法を利用することで、完全データで評価値が低い群を判断した場合と統合データで判断した場合とで、結果の解釈が全く逆になる場合があるとわかった。このことから、企業のメーカーは不完全データも使用して解析を行うことが望まれる。

本研究では、評価値の低い群が変化する事例を示したものの、今後の課題として、より網羅的な状況でのシミュレーションを行う必要があると考える。

参考文献

- [1] 楽天市場：<http://www.rakuten.co.jp/>.
- [2] 池田和史，服部元，松本一則，小野智弘，東野輝夫（2012）：マーケット分析のためのTwitter投稿者プロフィール推定手法，情報処理学会論文誌コンシューマ・デバイス&システム，Vol.2，No.1，pp.82-93.
- [3] 岩崎学（2002）：『不完全データの統計解析』，エコノミスト社，pp.7-9,173-174.

表4. 各製品における閾値を変化させたときの適合率と再現率の推移

閾値	製品A				製品B				製品C				製品D			
	適合率P		再現率R		適合率P		再現率R		適合率P		再現率R		適合率P		再現率R	
	男性	女性	男性	女性	男性	女性	男性	女性	男性	女性	男性	女性	男性	女性	男性	女性
0.0	84.5%	62.3%	85.8%	60.0%	75.3%	73.6%	76.1%	72.9%	79.4%	71.6%	79.4%	71.6%	74.6%	78.5%	73.6%	79.4%
0.5	88.2%	68.8%	73.9%	46.0%	80.2%	79.9%	63.6%	61.1%	88.3%	76.3%	67.8%	51.6%	77.0%	80.9%	60.4%	62.9%
1.0	91.6%	74.3%	55.8%	36.3%	82.0%	83.3%	43.3%	48.2%	91.3%	84.6%	53.7%	39.1%	79.1%	87.4%	47.2%	47.4%
1.5	93.1%	82.6%	31.9%	26.5%	86.8%	86.6%	30.2%	36.8%	93.9%	88.9%	29.6%	28.4%	83.6%	92.0%	35.4%	39.4%
2.0	96.5%	92.1%	19.9%	16.3%	89.0%	87.5%	21.3%	30.0%	96.9%	90.6%	20.3%	21.3%	90.7%	94.0%	27.1%	26.9%
2.5	97.3%	92.0%	13.0%	10.7%	88.7%	92.8%	15.4%	22.9%	100.0%	97.6%	12.5%	18.2%	95.5%	100.0%	14.6%	20.6%
3.0	100.0%	95.0%	9.3%	8.8%	100.0%	96.2%	12.1%	18.2%	100.0%	100.0%	9.0%	12.9%	100.0%	100.0%	9.7%	16.0%