

SVMによる商品レビューの評価値の根拠となる語の抽出 Extraction of words explaining user rating for product by SVM

清瀬 太一朗[†] 湯本 高行[†] 新居 学[†] 上浦 尚武[†]
Taichiro Kiyose Takayuki Yumoto Manabu Nii Naotake Kamiura

1. はじめに

近年、オンライン上の通販サービスが多く利用されている。通販サイト Amazon.co.jp では、5段階の星の数で表される評価値が付与されたレビューが投稿されている。この評価値はテキストの内容を踏まえた上での総合的な評価として付与されることが多いため、テキストに記された感想や批評などと大いに関係があると考えられる。

例として、「デザインが素晴らしく、機能も最高の商品である」という星5のレビューでは、商品の「デザイン」が「素晴らしい」こと、商品の「機能」が「最高」であることが評価値5の理由だと推測できる。よって本研究では、サポートベクターマシン（以下、SVM）[1]を用いたレビューの解析により、評価値の根拠となると考えられる、商品についてのある側面を表す「属性」と、商品についての感想を表す「意見」に分類される語の抽出を目的とする。

本研究に関連した研究として、点数付きの商品レビューに対して回帰分析により、点数予測を行う研究[2]がある。この研究では、本についての英文レビューのみを用い、商品の評価の程度を、SVRにより多段階の指標として数値化している。こちらの研究では、SVRの回帰係数から評価決定に重要だと思われる素性が抽出できることが可能であると示されているが、あくまで評価予測がメインであり、素性の抽出結果についての定量的な評価はなされていない。

本研究では、日本語で記述されたレビューを使用するため用いる素性が異なり、あくまで評価に影響のある語の抽出を主としている。また、SVRだけでなくSVCも用いて両者の比較を行うとともに、フィルタリングによる精度向上も検討している。さらに、様々なジャンルの商品に適用し、商品の評価の高低により区別しているため、より評価に影響のある素性が抽出できると考えられる。

2. 提案手法

2.1 SVMによる語の抽出

SVMでの識別式は(1)式で表され、入力 x とパラメータ w , b によって属するクラス y を決定する。

$$y = \mathbf{w}^T \mathbf{x} + b \quad (1)$$

SVMでは(2)式の制約条件のもとで $\|\mathbf{w}\|^2/2$ を最大化することでパラメータの推定を行うが、ラグランジュの未定乗数法を用いることで、ラグランジュ未定乗数の最大化という問題に置き換えが可能である。

$$y_i((\mathbf{w}^T \mathbf{x}_i) + b) \geq 1 \quad (2)$$

(2)式の y_i は $(\mathbf{w}^T \mathbf{x}_i) + b \geq 1$ のとき1, $(\mathbf{w}^T \mathbf{x}_i) + b \leq -1$ のとき-1となることを示す。これより、 w はラグランジュ未定乗数 a_i は \mathbf{x}_i がサポートベクトルではないときに0となることから、サポートベクトル \mathbf{x}_i^* によって決まり、(3)式で表される。

$$\mathbf{w} = \sum_{i=1}^N a_i y_i \mathbf{x}_i^* \quad (3)$$

\mathbf{x} の各次元を語に対応させた場合、この w が素性に用いた語の重みとなる。 w は識別関数を決定するパラメータであるため、この重みの大きさが評価に対する影響の大きさとなり、重みの値が正であれば好評、負であれば不評に寄与する。よって、この重みを指標とし、語の抽出を行う。

また、今回実験にはLIBSVM[3]を用い、カーネル関数には線形カーネルを使用する。一般的にSVMを使用する際にはRBFカーネルが良く用いられている。しかし、このようなカーネル法を用いて識別を行う場合、SVMは識別式ではなくカーネル関数を用いて識別を行う。その場合、 w は写像した空間でのベクトルについての重みとなり、素性との対応関係が得られないため、通常は w を求めることはない。しかし、本実験では(1)式における w から語の重みを得ることが目的であるため、実験には高次元への写像を行わない線形カーネルを使用する。さらに、クラス分類用のSVCと回帰分析用のSVRを使用し、両者の比較も行う。

SVCは二値分類であるため、5段階の評価値を好評と不評の2クラスに分けた上で分類を行う。好評と不評の分け方は、評価値ごとのレビュー数をもとにレビュー数の差が最も少なくなるようにクラス分けを行う。例えば、評価値5, 4を好評クラス、評価値3, 2, 1を不評クラスに分けたとき、好評クラスと不評クラスのレビュー数が同じならば、それを最良のクラス分けであるとする。

SVRとはSVMを回帰問題に拡張したものであり、回帰とカーネル関数を用いて予測モデルを作成する。SVRでは回帰係数から語の重みが得られる。

2.2 素性ベクトル

レビューを特徴量として扱うためにMeCab[4]を用いて形態素解析を行う。分割された形態素のうち、「良い」、「悪い」、「最高」、「最悪」、「味」、「デザイン」など物事の評価に関係ある形容詞や名詞を素性として用いる。

しかし、名詞でも「一」や「10」などの数詞、「番」や「枚」などの助数詞、「あれ」や「これ」などの代名詞、「さん」などの接尾や「こと」「はず」などの非自立など、単体で見たとときに意味が得られない語はあらかじめ除外する。また、MeCabの解析結果が未定義扱いの語も除外する。

また、素性の重み付けにはTF-IDFを使用する。TF-IDFは(4)式で表され、 $tf(t,d)$ はテキスト d 中の単語 t の出現頻度、 S はテキスト d 中出现する全ての単語の出現頻度の和、 $df(t)$ は単語 t が出現するテキスト数、 N はある商品に対するレビューの総数を表す。

$$\frac{tf(t,d)}{S} \times \log \frac{N}{df(t)} \quad (4)$$

[†] 兵庫県立大学, University of Hyogo

2.3 フィルタリング

根拠となる語を素性の重みの大きさのみで判断すると、商品の全レビュー中に 1 回しか出現しない語も含まれる可能性がある。しかし、商品のレビュー集合に対して、出現するレビュー数が極端に少ないような語はその商品の評価に対する影響が弱いと考えられる。また、レビューにおいて「良い」、「悪い」などの感想や意見を表す語に比べて、「メーカー」、「使用」など評価に関係はないが、多く記述される語が存在する。そのため、出現するレビュー数の少ない語と、レビューあたりに出現する割合が多い語（以下、出現割合）を除外する。本実験では SVM の識別式より得られる重みの大きい語から重要語を取ってくるため、得られた語に対してフィルタリングを行う。

3. 語の抽出実験

各データから評価値の根拠となると思われる語の抽出を行う。また、各レビューの解析結果から得られた語の重みの大きさ上位下位 20 件の抽出結果をもとに、抽出された語にどのような傾向があるのかを調査し、「意見」と「属性」に分類された語を正解とする。20 件中の意見および属性数を適合率とし、その値により評価を行う。

3.1 実験環境

本実験では通販サイトの Amazon.co.jp で取り扱われている商品のレビューを使用する。実験に用いるレビューは 20 ジャンルの商品を対象とし、商品ジャンルごとに高評価が多い商品と低評価が多い商品の 2 パターンを用いる。使用する商品のジャンルを表 1 に、高評価商品と低評価商品のレビュー数の分布をそれぞれ図 1、図 2 に示す

表 1 商品ジャンル

PC パーツ	照明
イヤホン	美容
カメラ	菓子
カー用品	食品
キッチン用品	飲料
ネット機器	園芸用品
ベビー用品	外付けメモリ
メモリカード	家電用品
健康器具	携帯アクセ
入力機器	洗剤

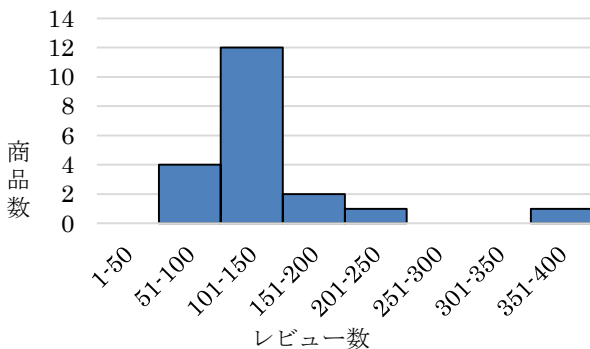


図 1 高評価商品のレビュー数内訳

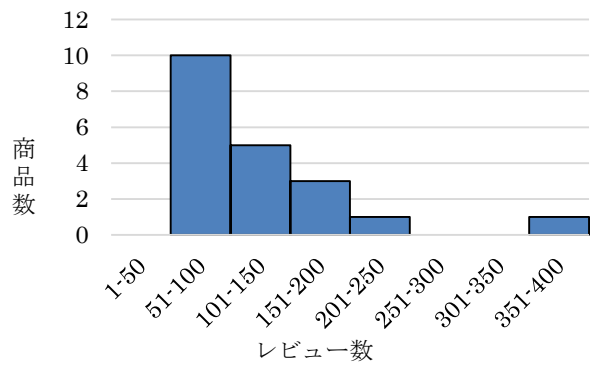


図 2 低評価商品のレビュー数内訳

また、高評価商品と低評価商品それぞれの評価値の内訳を割合として表したものを表 3、表 4 に示す。表 3、表 4 の棒グラフは評価値の平均値を、髭の上端は最大値を、下端は最小値を表している。

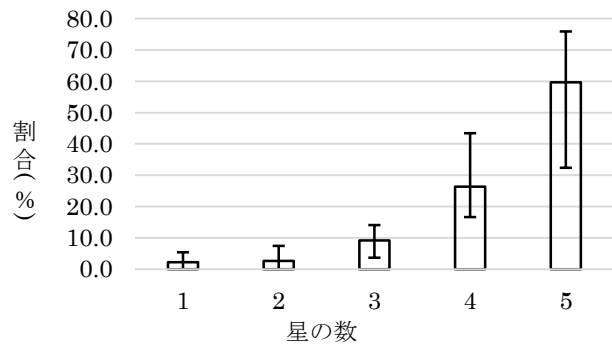


図 3 高評価商品レビューの評価値の内訳

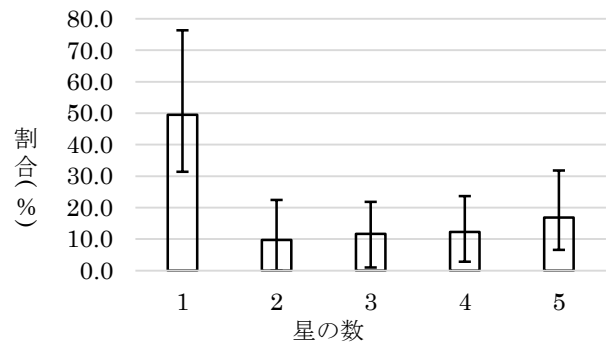


図 4 低評価商品レビューの評価値の内訳

3.2 抽出精度による評価

SVC と SVR を用いて語の抽出を行う。また、抽出を行う際にフィルタリングを用いた場合についても検討する。出現頻度によるフィルタリングでは、出現頻度が 1 回の語を除外する。出現割合によるフィルタリングはしきい値を変化させ最も結果のよい値を使用する。

SVC と SVR で出現割合によるフィルタリングのしきい値を変化させ、重みの上位下位 20 件中の意見および属性

の数を適合率とし、その値により評価を行う。出現割合によるフィルタリングを使用しない場合としきい値を変化させた場合での適合率の変化をそれぞれ図 5、図 6 に示す。このとき、出現頻度によるフィルタリングも同時に行っている。

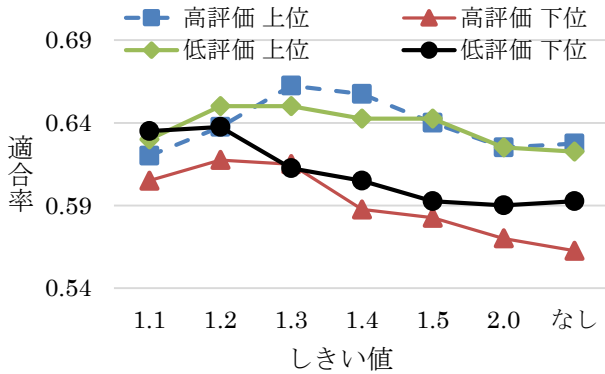


図 5 SVCでの出現割合のしきい値変化

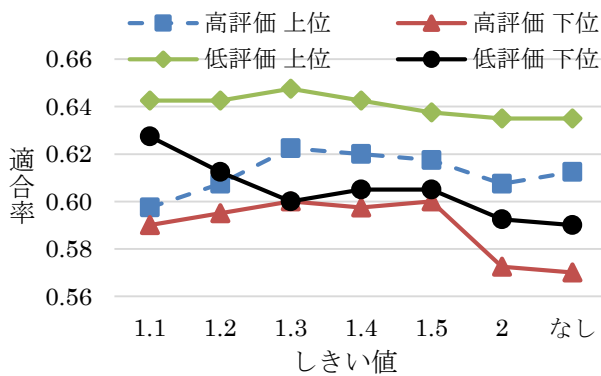


図 6 SVRでの出現割合のしきい値変化

以上の結果より、出現頻度と出現割合のフィルタリングを組み合わせると効果的であることが分かる。また、出現割合のしきい値は SVC では 1.2, SVR では 1.3 としたとき最も多くの意見と属性が抽出できることが分かった。

以上を踏まえて、SVC と SVR による抽出結果の比較とフィルタリングの有無による結果の比較を行う。それぞれの場合での適合率を表 2 に示す。

表 2 の結果よりフィルタリングを使用した方が総じて適合率が向上した。また、SVC を用いて、フィルタリングを行ったときに最も多くの意見と属性が抽出できることが分かった。

表 2 手法別の適合率

	フィルタ	高評価		低評価		平均
		上位	下位	上位	下位	
SVC	有り	0.64	0.62	0.65	0.64	0.64
	なし	0.61	0.55	0.62	0.59	0.59
SVR	有り	0.62	0.60	0.65	0.60	0.62
	なし	0.60	0.55	0.65	0.59	0.60

3.3 出現頻度に基づく語の抽出方法との比較

今回実験では、SVM を用いることで得られる語の重みを指標として、語を抽出しているが、単純に商品レビュー中の語の出現頻度に基づいて抽出を行うことも可能である。そこで、全商品に対して商品レビュー中に出現する語の出現頻度の多い順に上位 20 件の語を抽出し、意見と属性に分類できた数を適合率として表し、提案手法との比較を行う。抽出結果を表 3 に示す。

表 3 出現頻度に基づく抽出

フィルタ	高評価	低評価	平均
なし	0.50	0.52	0.51
有り	0.56	0.57	0.56

表 3 の結果より、出現頻度にもとづく語の抽出と比較して提案手法の方がより多くの意見と属性を抽出できることが分かる。また、重みの大きさをもとに抽出を行う場合、重みの値が正か負かによって、語が高評価と低評価のどちらかに影響しているか判別することも可能だが、出現頻度の多さのみに着目した場合には、それが不可能である。以上のことより出現頻度のみにもとづいて語の抽出を行うよりも、提案手法の方が優れていることが分かる。

3.4 考察

得られた意見と属性のうち抽出数の多かった語各 5 件を表 4 に示す。

表 4 抽出された語

高評価		低評価	
上位	下位	上位	下位
満足	残念	満足	返品
値段	悪い	よい	使い物
簡単	大きい	問題	残念
やすい	普通	便利	悪い
早い	レビュー	無い	最悪

表 4 では、高評価の上位では「満足」「簡単」などの意見が、低評価の下位では「残念」「最悪」などの意見が抽出できているため、高評価の商品に対する好評な意見や、低評価の商品に対する不評な意見を得ることが可能である。しかし、高評価の下位にある「レビュー」という語は意見でも属性でもない。このような語が適合率向上の妨げになっていると考えられる。

また、抽出された語について、意見と属性にわけて考える。得られた意見と属性の内訳を表 5 に示す。

表 5 結果中の意見・属性の割合

	高評価		低評価	
	上位	下位	上位	下位
意見	0.53	0.45	0.54	0.52
属性	0.11	0.17	0.11	0.12

表 5 より、抽出対象の 20 件中の約半分が意見であり、正解として得られた抽出結果の約 8 割が意見であることがわかる。意見が多く得られると、商品がどう良いのかまたはどう悪いのか、というものが分かりやすくなると考えられる。しかし、属性の方は 20 件中から 1 割程度しか得ら

れていないことがわかる。そのため、商品の何が良いのか何が悪いとされているかが把握しづらいと考えられるため、属性の抽出結果に絞って検討を行う。「良い」や「悪い」などの意見は、商品のジャンルに関係なく多くのレビューに出現しやすいが、属性は商品のジャンルによって固有のものが多い。また、表5に示したように属性は意見に比べ大幅に少ない。そのため、表4のように抽出数が多い語のみでは、意見しか確認ができないと考えられる。そのため、得られた属性をジャンルごとにまとめ、ジャンルごとに得られた属性について検討する。ジャンルごとの属性の一部を表6に示す。

表6 各ジャンル別の属性

ジャンル	属性
PC パーツ	音, 品質, サポート, コスト
イヤホン	体質, 具合, 見た目, 接触
カメラ	サイズ, 重量, 値段, 画質
カー用品	機能, 出力, 値段, 説明
キッチン用品	味, 使い勝手, タイプ, 使い方
ネット機器	品質, 性能, 価格, 相性
ベビー用品	枚数, 値段, 素材, 使い勝手
メモリカード	変換, 容量, スピード, 耐久
健康器具	値段, 機能, 消耗, 音
入力機器	デザイン, 価格, 形状, レスポンス
園芸用品	性能, 耐久, 見た目, 繋ぎ
外付けメモリ	容量, 価格, 読み込み, 接触
家電用品	価格, 性能, 容量, 値段
携帯アクセ	外観, デザイン, 表面, 画面
洗剤	仕様, 効果, 分量, 使い勝手
照明	明かり, 用途, 機能, 熱
美容	価格, サポート, サイズ, 繊維
菓子	魅力, 状態, 品質, 衛生
食品	辛み, 味, 衛生, 旨み
飲料	風味, カロリー, 価格, 甘味

表6より、販売商品のレビューであるため、ほぼ全てのジャンルにおいて、「値段」または「価格」といった語が存在していることが分かる。他には、PC パーツやネット機器に対しての「品質」や、メモリカードや外付けメモリに対しての「容量」といった、その商品を特徴付ける語や関連する語が得られていることがわかる。よって、抽出数は少ないものの、ジャンル特有の語は抽出できていることが分かる。

また、本研究は語の抽出が主であるが、SVMによる分類の精度についても検討を行う。SVCでは分類の正答率、SVRでは回帰分析の平均二乗誤差により評価を行う。SVCでの正答率を図7に、SVRの平均二乗誤差を図8に示す。

図7よりSVCでは総じて5割以上の正答率であり、低評価の商品の方が正答率は高くなっている。また、図8では、高評価商品の平均二乗誤差はほぼ1以下であり、最小で0.35となっている。しかし、低評価の商品では全ての商品において平均二乗誤差が1以上であり、最大では2.80となっている。図4の低評価商品の評価値の内訳を見ると、評価値1の次に評価値5が多いため、両極の評価値に予測が引っ張られることで誤差が大きくなってしまったと考えられる。しかし、平均平方根二乗誤差は最大で1.67であり、

全体を平均しても1.38であるため、極端に精度が悪い訳ではない。

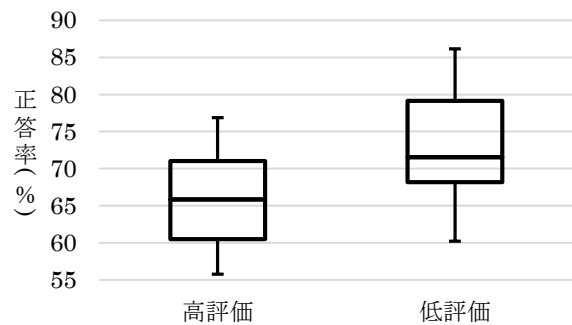


図7 SVCによる分類精度

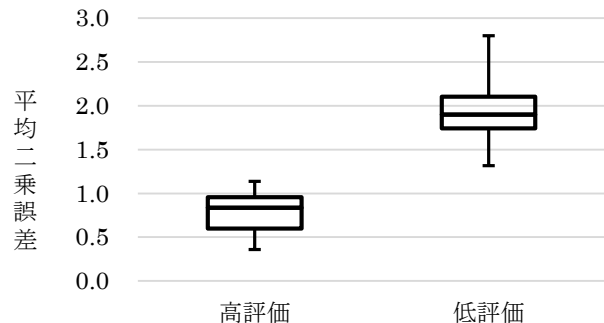


図8 SVRによる分類精度

4. おわりに

SVMによって商品レビューの評価値を語に基づいて予測するモデルを構築し、モデルを用いて根拠となる語を抽出する手法を提案した。SVCによる分類とSVRによる回帰分析を用いて、提案した出現頻度と出現割合という2つの指標をもとにしたフィルタリングを用いてそれぞれの適合率を比較した結果、SVCを用いて出現頻度1、出現割合1.2をしきい値とした場合で適合率0.64と最大となった。抽出結果を確認すると、表2にあるように「レビュー」のような評価に関係なく出現する語が多く抽出されてしまっている。また、ジャンル特有の語が抽出できてはいるものの、表5にあるように属性数が著しく少ないことが上げられる。

以上より、どのように不要な語の排除を行いながら、適合率および属性数を向上させるかが今後の課題である。

謝辞

本研究の一部は、平成27年度科研費若手研究(B)「情報の詳細関係に基づくWebページの組織化」(課題番号:24700097)によるものである。

参考文献

- [1] C Cortes, V Vapnik, "Support Vector Networks", Machine Learning, 20, pp.273-297 (1995).
- [2] 岡野原 大輔, 辻井 潤一, "レビューに対する評価指標の自動化と", 自然言語処理, Vol.14, No.3, pp.273-295 (2007)
- [3] LIBSVM, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- [4] MeCab, <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>