複数の評価項目を持つレビューの評価値の推定

嶋田 和孝　小濱 祐樹　遠藤 勉
九州工業大学 情報工学部 知能情報工学科
{shimada, y_kohama, endo}@pluto.ai.kyutech.ac.jp

1　はじめに

近年の WWW の普及により，膨大な文書が蓄積されている．日々増え続けていくこのような文書に対して，的確な情報抽出や分類・要約技術の重要性が増している．現在，我々は Web から，大量のレビュー記事を入手する事ができる．そして，その記事を製品購入などの意思決定の際に参考にしている．しかし，大量のレビュー記事の内容を把握するのは容易ではなく，内容を把握するために，記事の効率的な活用法が必要となる．その結果，近年， Web 上に存在する評判情報の抽出や分類に関する研究が数多く行われている [6].

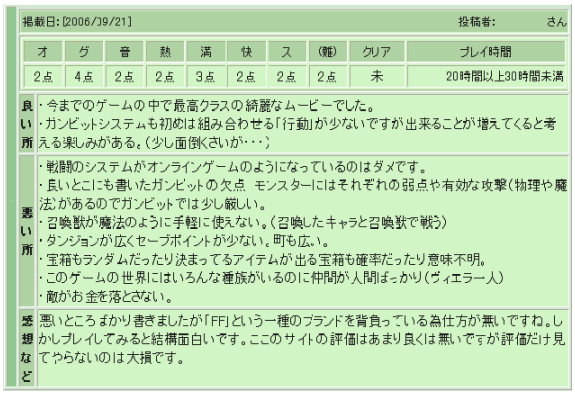
評価表現に対する研究は，評価表現の抽出や辞書の構築，評価文書の分類など様々なタスクが存在するが，本研究では，評価文書の極性判断をその研究の対象とする．従来の評価文書の分類はレビュー記事を肯定的 (p)/ 否定的意見 (n) に自動分類する p/n 分類を対象とするものが多かった [3, 5]．一方で，近年， p か n かの 2 値ではなく，より粒度の細かい評価値の分類をタスクとする研究がなされている [2, 8]．本論文で扱うタスクは，これらと同様に 2 値ではなく，より粒度の細かい分類を対象とする．

さらに，従来の研究では，レビューの極性は 2 値である場合も多値である場合も 1 文書につき 1 つであり， 1 つの文書が複数の評価項目とその評価値を保持している場合を考慮していない．しかしながら，例えば，ある製品や映画などのレビューを考えると，最終的にはその対象についての全体的な善し悪しを p/n もしくは 5 段階などで表すことが多いが，実際には評価者はより細部についても評価しているはずである． PC を対象とした場合は，性能や操作性，携帯性など，映画の場合は脚本，演出，映像，音楽などである．ここでは，このように 1 つのレビュー記事が複数の項目を保持し，それぞれの評価値を分類・推定することを多項目分類と呼ぶことにする．製品全体の評価のみではなく，その細部まで評価することは，そのような評価情報を閲覧するユーザにとって有用である．さらに，レビュー中のどのような語や要素がどの項目に対して有効な指標となるのかを分析することは，評価表現を扱うタスクとして極めて重要である．

本研究では， Web 上に存在する多項目で評価されたレビュー記事を対象とし，それらの項目の評価値を推定する枠組みについて検証する．この多項目・多値分類のタスクに対して， SVR を評価値推定の手法として，いくつかのパラメータについて比較実験を行い，考察する．

2　タスク

Web 上には様々な製品のレビュー記事が存在するが，本論文では，あるゲーム機のソフトウェアのレビュー記事を処理の対象とする．図 1にレビュー記事の例を示す．このレビュー記事では， 1 つの記事に「良い所」，「悪い所」，「感想」の 3 つの記述欄があり，評価項目として「オリジナリティ (オ)」，「グラフィックス (グ)」，「音楽 (音)」，「熱中度 (熱)」，「満足度 (満)」，「快適度 (快)」，「ストーリー (ス)」，「難易度 (難)」の 8 つがある．評価値のスケールは 0 ～ 5 までの 6 値である．


掲載日:[2006/09/21]　投稿者:　さん

オ	グ	音	熱	満	快	ス	(難)	クリア	プレイ時間
2点	4点	2点	2点	3点	2点	2点	2点	未	20時間以上30時間未満

良い所
・今までのゲームの中で最高クラスの綺麗なムービーでした。
・ガンビットシステムも初めは組み合わせる「行動」が少ないですが出来ることが増えてくると考える楽しみがある。(少し面倒くさいが・・・)

悪い所
・戦闘のシステムがオンラインゲームのようになっているのはダメです。
・良いとこにも書いたガンビットの欠点 モンスターにはそれぞれの弱点や有効な攻撃(物理や魔法)があるのでガンビットでは少し厳しい。
・召喚獣が魔法のように手軽に使えない。(召喚したキャラと召喚獣で戦う)
・テンションが広くセーブポイントが少ない。町も広い。
・宝箱もランダムだったり決まってるアイテムが出る宝箱も確率だったり意味不明。
・このゲームの世界にはいろんな種族がいるのに仲間が人間ばっかり(ヴィエラー人)
・敵がお金を落とさない。

感想など
悪いところばかり書きましたが「FF」という一種のブランドを背負っている為仕方が無いですね。しかしプレイしてみると結構面白いです。ここのサイトの評価はあまり良くは無いですが評価だけ見てやらないのは大損です。

図 1: レビュー記事の例

本論文で使用するレビュー記事は 2 つの Web サイトから抽出した[1,2]．これらのサイトでは，レビューの投稿に際してのガイドラインがあり，投稿されたレビューは人手でチェックされ，条件に沿わない投稿記事は Web サイト上に掲載されないシステムになっており，比較的良質なデータであると考えられる．

3　手法

3.1　SVR

このタスクで，各項目の評価値を推定する手法として Support Vector Regression(SVR) を用いる． SVR は SVM の回帰問題への拡張であり，先行研究でも評価値の多値分類に使用されている．先行研究と同様に SVM を多クラス問題へ拡張することも考えられるが，予備実験において one-versus-one 法に基づく SVM と SVR を比較した結果，正しい評価値と分類器の出力の平均二乗誤差で比べた場合， SVR と SVM には明確な差があったので (平均二乗誤差で 0.2 程度)，今回は SVR のみを実験対象とする． Pang ら [2] は Metric Labeling を SVM や SVR と組み合わせることで，正解率が向上することを示しているが，本論文では SVR 単体で評価する．

使用する素性は，レビュー中の「良い所」と「悪い所」に書かれているテキストから抽出された単語群である．「感想」部分を使用しないのは，予備実験において「感想」を含まない場合の方が精度が高かったためである．

[1] http://psmk2.net/
[2] http://ndsmk2.net/

素性空間では，同じ単語でも「良い所」に書かれた単語と「悪い所」に書かれた単語は別のものとして扱う．すなわち，ある単語 w_i について，「良い所」に書かれた単語は w_i^p，「悪い所」に書かれた単語は w_i^n として扱い，あるレビュー記事 d_x の項目 y における素性は

$$d_{xy} = \{w_1^p, w_2^p, ..., w_j^p, w_1^n, w_2^n, ..., w_j^n\}$$

のようになる．ここで j はレビュー中で素性として使用される単語の総数である．本論文では，素性として使うのは形態素解析の結果の品詞が「名詞」，「形容詞」，「副詞」のどれかであるものである．素性の値は，その単語のテキスト中での有無によって決定され，頻度情報などは用いていない．

3.2 素性選択

高精度な分類器を生成するには有効な素性の選別が必要になる．本研究で扱うような 1 つのレビューが複数の評価項目を含む場合，レビュー中の全ての語が全ての評価項目に関連しているとは考えづらい．すなわち，各項目ごとにその項目の評価を支えている語が存在するはずである．そのような素性を抽出するために，素性候補となる全ての語に対して，各項目の各評価の値を基にその語の信頼度を算出する．ここで信頼度とは，ある評価値におけるその単語の分散だと考える．すなわち，各単語に対して，ある評価項目でどれだけ同じ点数でその語が出現するか，というものを信頼度だとする．

$$var(w_{c_j}) = \frac{1}{m} \sum_{i=0, w \in d_i}^{n} (real(d_i, c_j) - ave(w_{c_j}))^2 \quad (1)$$

ここで c_j はある評価項目であり，m は単語 w の文書頻度 (df 値) に相当する．n は文書の総数である．ここで，文書 d_i に単語 w が存在しない場合は計算の対象から外す．$real(d_i, c_j)$ はレビュー記事 d_i における c_j の実際の評価値を表し，$ave(w_{c_j})$ は w の c_j における評価値の平均である．この var がある閾値以下のものを素性として使用する．

また実験では，その他にもいくつかの条件を比較考察する．

頻度 (F) 単語 w の出現頻度が n 以上のもの．

評価値 (E) 単語 w が「良い所」に生じた単語である場合，その評価値の平均が 3 以上であること．「悪い所」に生じた単語である場合，その評価値の平均が 3 以下であること．

4 実験

4.1 データセットと評価基準

2 種類のデータセットを用意し，その有効性を検証した．1 つめのデータセットはある 1 つのソフトのレビュー記事のみを対象としたものであり，レビュー数は 553 記事である．2 つめのデータセットは，あるゲーム機に対する複数のソフトのレビューが混在したものであり，レビュー数は 1114 記事である．実験では，これらのデータについて交差検定 (Leave-one-out) を行った．実験結果は，SVR の出力と実際の評価値の平均二乗誤差 (MSE)，適合率 (P)，および再現率 (R) で評価した．平均二乗誤差 (MSE) は以下のように算出される．

$$MSE_j = \frac{1}{n} \sum_{i=1}^{n} (svr(d_{ij}) - real(d_{ij}))^2 \quad (2)$$

ここで i はレビュー記事，j は評価項目を指す．svr と $real$ はそれぞれ SVR の出力結果と実際の評価値である．但し，ここで，SVR の値は実数ではなく，四捨五入することによって整数化されている．適合率 (P)，および再現率 (R) は以下のように算出される．

$$P = \frac{\text{正しく評価値を推定できた項目の数}}{SVRが評価値を算出できた項目の数} \quad (3)$$

$$R = \frac{\text{正しく評価値を推定できた項目の数}}{\text{評価項目の総数}} \quad (4)$$

また，素性選択の基準によっては，あるレビューのある項目に対して素性そのものが存在しない場合が出てくる．そのため，実験結果にはどの程度の項目を扱えたかを表すカバー率 (C) も併記する．

$$C = \frac{SVRが評価値を算出できた項目の数}{\text{評価項目の総数}} \quad (5)$$

4.2 実験結果

まず初めに，1 つのソフトのみのレビュー記事を対象とした場合の実験について述べる．ここでは，サイト内で最も投稿レビュー数が多かった Final Fantasy XII を実験対象とした．

最初に，素性選択において，単語の評価値の分散により，素性を制限しなかった場合の SVR とベースラインとの比較をする．実験結果を表 1を示す．表中の All-3 は全ての項目が 3 点だと仮定した場合の平均二乗誤差であり，Ave は全データから各項目の評価値の平均を求め，それらをそれぞれの項目の点数と考えた場合の平均二乗誤差であり[3]，これらは一つのベースラインとなる．表中の SVR の項目欄で，F の値は素性選択における単語頻度の条件で，E は素性選択の節で説明した，語の出現場所における評価値の条件を用いた場合の素性であることを表している．上記 2 つの指定がない SVR は品詞制限のみの素性群である．表から分かるように，SVR はどの場合でもベースラインより正しく評価値を推定していることが分かる．素性選択において var の値を利用しない場合は，どの場合もカバー率は 1 であり，この場合，再現率と適合率は同値になり，平均二乗誤差のような大きな差は見られなかった．また，頻度情報や評価値情報も有効に機能しなかった．

[3] この場合も得られた平均値は整数化して平均二乗誤差を計算している．

表 1: 実験結果: ベースラインとの比較

		All-3	Ave	SVR	SVR F ≥ 4	SVR E	SVR F ≥ 4 & E
MSE	オリジナリティ	1.15	1.15	0.87	**0.84**	0.99	1.01
	グラフィックス	2.11	0.62	**0.49**	0.51	0.52	0.57
	音楽	1.30	1.30	0.77	**0.76**	0.79	0.78
	熱中度	1.86	1.86	**1.01**	1.05	1.09	1.12
	満足度	2.32	2.04	**0.95**	**0.95**	0.99	0.97
	快適度	1.63	1.63	**1.07**	1.11	1.08	1.14
	ストーリー	2.98	1.43	**0.78**	0.81	0.87	0.91
	難易度	1.00	0.66	**0.65**	**0.65**	0.65	0.66
	平均	1.79	1.34	**0.82**	0.84	0.87	0.90
適合率 (P)		0.23	0.34	0.45	0.45	0.45	0.44
再現率 (R)		0.23	0.34	0.45	0.45	0.45	0.44
カバー率 (C)		1	1	1	1	1	1

表 2: 実験結果: 素性選択 (SVR, var=0.75)

		A	B	C	D
MSE	オリジナリティ	0.89	0.76	0.91	**0.74**
	グラフィックス	0.54	0.56	**0.52**	**0.52**
	音楽	0.90	0.76	0.82	**0.75**
	熱中度	1.08	0.87	1.04	**0.85**
	満足度	1.24	0.83	1.19	**0.71**
	快適度	0.99	0.88	0.99	**0.79**
	ストーリー	0.91	0.77	0.96	**0.62**
	難易度	**0.62**	0.65	0.63	0.71
	平均	0.90	0.76	0.88	**0.71**
適合率 (P)		0.46	0.44	0.44	0.44
再現率 (R)		0.45	0.39	0.42	0.36
カバー率 (C)		0.99	0.88	0.97	0.83

表 3: 実験結果: 素性選択 (SVR, var=0.5)

		A	B	C	D
MSE	オリジナリティ	0.91	**0.51**	0.88	0.48
	グラフィックス	0.52	**0.43**	0.53	0.45
	音楽	0.94	0.61	0.86	**0.50**
	熱中度	0.98	**0.47**	0.95	0.50
	満足度	1.38	**0.36**	1.31	**0.36**
	快適度	0.87	0.46	0.85	**0.38**
	ストーリー	0.99	0.55	0.98	**0.32**
	難易度	**0.57**	0.64	0.63	**0.57**
	平均	0.90	0.50	0.87	**0.45**
適合率 (P)		0.47	0.48	0.46	0.49
再現率 (R)		0.46	0.35	0.43	0.33
カバー率 (C)		0.97	0.72	0.95	0.66

次に var を用いて素性選択を行った場合についての実験結果を比較する．var の値が 0.75 以下の語を利用した場合と var が 0.5 以下の場合の結果をそれぞれ表 2と表 3に示す．両方の表において

A SVR: 条件は var の値のみ

B SVR: 条件は var の値， F≥4

C SVR: 条件は var の値，条件 E

D SVR: 条件は var の値， F≥4，条件 E

を表す．var の値を素性選択に用いた場合は，用いなかった場合と異なり，頻度情報や出現位置による評価値の条件の 2 つを利用した方が再現率およびカバー率は下がるものの平均二乗誤差ではよい値を得た．特に頻度情報が効果を示すことが実験結果から分かる．一方で，評価値の条件のみの場合，カバー率が低下しているにもかかわらず，その正確さは var を利用しなかった場合よりも悪くなる傾向があり，今後考察が必要である．

本タスクでは，1 つのレビュー記事が複数の評価項目とその評価値を持っており，実際には，例え人間がそのテキストを見ても，そのレビューのみでは全ての評価値を推定できないという可能性がある．すなわち，レビューによっては，本質的に，評価値を推定できないという問題が生じる．これは，実験の際にどの程度までカバー率の低下を許容するか，という問題に繋がる．そこで，実験に使用した 553 記事からランダムに選択した 30 記事について，そのテキスト部分から各項目の値を推測できるか，著者が主観的に評価した．但し，ここで，「推測できるか」という基準は，具体的な数値を人間が特定できるかというものではなく，例えば，ある項目の評価値が 4 点や 5 点の場合，レビューの「良い所」の欄にあるテキストを読むことで，その項目を肯定的に評価していることが推測できるか，というレベルで判断した．その結果，全項目のうち 75% 程度が推測可能であると判断された．この結果を踏まえると，var=0.75 における B や D のカバー率でも許容範囲であり，素性選択はある程度の効果があることが推測される．しかし，今回の推測可能性に関する評価は著者一人で行っており，複数の被験者による揺れがどの程度生じるのか，また，人間が推測できないと判断した項目と学習器が扱うことのできなかった項目の一致度がどの程度あるのかなどについて詳しい考察が必要である．

最後に異なるソフトのレビューが混在したデータセットについての実験について述べる．レビューは NintendoDS のソフトについて，RPG やアクションなど，さまざまなジャンルから選ばれている．実験結果を表 4に示す．表中の SVR1 ～ SVR3 はそれぞれ SVR1) 素性選択は品詞のみ，SVR2) var が 0.75 以下，頻度 4 以上，条

表 4: 実験結果: 異なるソフトのレビュー

		Ave	SVR1	SVR2	SVR3
MSE	オリジナリティ	1.20	0.74	0.79	**0.57**
	グラフィックス	0.85	0.64	0.63	**0.57**
	音楽	0.71	0.79	0.69	**0.64**
	熱中度	1.71	0.91	0.90	**0.44**
	満足度	1.76	1.06	0.84	**0.46**
	快適度	1.29	1.01	0.84	**0.44**
	ストーリー	2	0.23	0.27	**0.19**
	難易度	1.17	0.91	0.58	**0.24**
	平均	1.34	0.76	0.69	**0.44**
適合率 (P)		0.38	0.45	0.43	0.47
再現率 (R)		0.38	0.45	0.38	0.33
カバー率 (C)		1	1	0.88	0.71

件 E, SVR3) *var* が 0.5 以下，頻度 4 以上，条件 E を満たすものである．タスクとしては，1 つの製品に対するレビューだけというタスクよりは，複数の製品が混在した場合の方が問題としては難しくなると考えられる．実験データ数が FFXII の場合と異なり，2 倍近くあるため単純に比較はできないが，ベースラインと比べ，平均二乗誤差やその他の値が改善していることを考えると，比較的良好な結果が得られていると判断できる．

5 考察

実験結果を踏まえて考察する．いくつかの指標に基づき素性選択を行った結果，適合率や再現率は十分とは言えないが，ある程度，評価値の推定が可能であることが分かった．今回は，素性に単語しか用いていないが，素性の種類については議論が必要である．本タスクの場合，「オリジナリティ」や「グラフィック」といった特定の評価項目について，その評価値を推定するという特性があるため，n-gram や部分依存木といった素性を組み込むことで，例えば「グラフィックが良い」のような表現を効率的に学習に活用でき，大きな精度向上に繋がると考えられる．また，上記の特性を利用して，記事全体を推定に使うのではなく，例えばレビュー中で「オリジナリティ」や「グラフィックス」という語と関連度の高い語を選び出し，その語とそれに関連する評価表現のみを用いて評価値を推定する方が効果的な可能性もあり，小林ら [7] が提案しているような評価要素の抽出タスクの結果を利用することができるかもしれない．

今回はデータ数が少なかったため，交差検定でその精度を評価したが，Goldberg ら [1] が主張しているように，訓練データの獲得も大きな問題である．筆者らはタスクは異なるが，類似度に基づく訓練データの獲得について，それが分類精度の向上に繋がることを確認しており [4]，少ない訓練データをシードとし，良質な訓練データを自動獲得する手法についても考察が必要となる．

今回は SVR のみを用いて，評価実験を行ったが，同様の評価文書の多値分類のタスクで Metric Labeling を適用することで精度が向上することが知られている [2]．また，堤ら [9] は対象は p/n 分類であるが，複数の分類器を利用して，その結果を基に各分類器の出力結果の信頼度を算出し，組み合わせることで，分類精度が向上することを示している．本タスクにおいても，新たな評価値推定手法について今後議論が必要である．

6 おわりに

本論文では 1 つのレビュー記事が複数の評価項目を持ち，さらに p/n ではなく，より粒度の細かい評価値を持つようなタスクにおいて，SVR を用いて，その評価値を推定可能かどうか実験的に検証した．素性選択の手法にはまだ改善の余地があり，学習器についてもさらなる議論が必要である．

参考文献

[1] A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*, 2006.

[2] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 115–124, 2005.

[3] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, 2002.

[4] K. Shimada and T. Endo. Acquisition of new training data from unlabeled data for product specifications extraction. In *Proceedings of PACLING 2005*, pp. 284–289, 2005.

[5] P. D. Turney. Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[6] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–242, 2006.

[7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 2, pp. 203–222, 2005.

[8] 岡野原大輔, 辻井潤一. 評価文に対する二極指標の自動付与. 言語処理学会第 11 回年次大会, 2005.

[9] 堤公孝, 嶋田和孝, 遠藤勉. 複数の分類結果の信頼度を利用したレビュー記事の自動分類. 人工知能学会第 63 回 人工知能基礎問題研究会, SIG-FPAI-A601, pp. 27–32, 2006.