2B1-05

電子掲示板上の評判情報に基づく意思決定支援

Decision Support Based on Reputation from Bulletin Board

藤村　滋　Shigeru FUJIMURA
松村　真宏　Naohiro MATSUMURA
岡崎　直観　Naoaki OKAZAKI
石塚　満　Mitsuru ISHIZUKA

東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology,The University of Tokyo

Abstract:This paper describes a method of extracting reputation from bulletin board based on machine learning(C4.5), and shows that the analysis of decision trees is useful for decision support in new product development or marketing.

1. はじめに

電子掲示板上の評判情報を新製品開発やマーケティング戦略のための意思決定に応用することが注目を集めている．実際，人々の意見が集まる電子掲示板などは，企業の担当者によってチェックされている．

しかし，日々変化しつづけている Web 上の膨大な情報の前では，人手によるチェックなど波打ち際で砂の城を築くようなものである．ヒューリスティックな知識を用いて評判情報を抽出する試みもなされているが，そのような知識は対象ごとに作成しなければならず効率的でない．そこで本稿では，コールセンターのログなどのユーザの意見がタグ付けされたコーパスから，評判情報を抽出するためのルールを機械学習 (C4.5) により発見する手法を提案する．

2. 機械学習によるルールの発見

2.1 機械学習による評判情報抽出

掲示板のコメント一つ一つを評判 (良)，評判 (悪)，その他に分類することで，評判情報を抽出する．さらに，分類するために用いられたルールを分析し評判の原因を製品開発等の意思決定に利用することを考える．

テキストの分類に対して，従来から機械学習はよく用いられてきた．本稿では C4.5[Quinlan 95] を用いる手法を提案する．C4.5 は決定木学習の一つであり，生成されたルールが人間にも理解しやすいというメリットがある．生成されたルールである決定木を評判の原因の分析に利用するため，ここでは C4.5 を用いることにした．

C4.5 に与える属性と属性値は，以下のようにした．

- （属性）形容詞，$tf \cdot idf$ 法による重要単語
  (名詞，動詞，副詞，未知語)，
  類語検索による属性拡張，
  助動詞「たい」「ない」，
  接頭辞　「非・不・無・未」 ＋ 名詞
- （属性値）出現回数 (zero,more than one)

属性の説明について，形容詞は日本語で対象への評価を表す際に用いられる品詞であるので，評判情報においては最も重要な属性である．$tf \cdot idf$ 法で得られた重要単語を属性として用いたのは，例えばメーカー名・ブランド名のような言葉が決定木に取り込まれることによって，評判の原因分析に役立つと考えたからである．類語検索による属性拡張では，どの分野でも評判に直結する「良い」「悪い」を拡張することで，名詞「最高」などにも対応させた．助動詞については,「ない」は用言と結びついて意味を反転させるので重要な属性であり,「たい」は「〜を買いたい」のように評判を表す可能性を考慮し属性として採用した．接頭辞＋名詞は，例えば「未解決」「非対応」といった言葉に対応させるため属性に加えた．また，属性値にいてはそのキーワードが出現するかしないかの二値とした．

2.2 従来手法との比較

従来からの評判抽出手法としては，キーワードを用いたヒューリスティックルールによる手法がある．ヒューリスティックを応用するという手法の例として，立石ら [立石 01] は，キーワードを集めた評価表現辞書のみでなく，評価表現の現れる近辺を意見とする近接演算処理，および構文解析による適正値判定処理を行うことにより Web 上の評判情報検索を行い，検索結果の上位 17.1%で適合率 86.6%，また全体でも適合率 48.0%，再現率 48.6%という結果を実現している．

本手法を従来からの評判抽出手法と比較すると，次の点が考えられる．従来手法では，再現率を向上させるにはキーワードを増やすことになるが，評価表現は多種多様であり巨大な評価表現辞書を作成することはコスト面でも大きな負担を要する．特に，製品ごとに固有なキーワードを考慮する必要性があるため，対象製品が多くなると非効率的である．

一方，本手法では形容詞，$tf \cdot idf$ による重要語を決定木の属性として加えたことで製品に固有なキーワードも自動的に組み込まれる．また，想定していなかった意外なキーワードが発見されれば，消費者の新しいニーズを探ることもできる．さらに，複雑なルールを人手で考えなくとも，C4.5 という一般的な手法で，手軽にルールを発見できるという利点がある．

3. 評価実験

本手法について，価格.com[*1] のノート PC に関する掲示板のあらかじめタグ付けされた 2850 コメントからなるコーパスを用いて，5 分割交差検定によって評判情報抽出の精度・再現率を求める評価実験を行った．また，評価のベースラインとして以下のような，簡単なヒューリスティックルールにより，キーワードの出現の有無によって評判情報を抽出した．

- いい，良い，素晴らしい，最高，良好，ベスト
  → 評判 (良)

連絡先: 藤村　滋，東京大学大学院情報理工学系研究科，fujimura@tkl.iis.u-tokyo.ac.jp

*1 http://www.kakaku.com

- 悪い，ひどい，最低，最悪，不良，粗末
  → 評判 (悪)
- 上記の単語が一度も出てこない → その他

それぞれ，結果は表 1，表 2 のようになった．

表 1: ヒューリスティックルールによる精度・再現率　単位 (%)

	評判 (良)	評判 (悪)	その他
Precision	61.5	50.8	46.4
Recall	45.5	27.4	77.4

表 2: 提案手法による精度・再現率　単位 (%)

	評判 (良)	評判 (悪)	その他
Precision	66.3	46.8	64.2
Recall	65.5	43.8	68.6

ヒューリスティックにおいては，曖昧なキーワードを増やすことによって精度が低下することを懸念し，高精度の比較対象とするため，文意を一意に決定しやすいキーワードのみを用いた．

提案手法において，大きく再現率が向上したことは考慮しているキーワード数が 40 倍以上違うことが原因である，一方，精度については評判 (良) では向上し，評判 (悪) においても 4%の低下に留まっている．評判 (良) で精度が向上した原因としては，ヒューリスティックルールにおいて「いい」が「〜するといい」のように，評判と関係なく用いられていることが多かったことが原因として考えられる．

4. 考察

4.1 決定木の分析による意思決定支援への応用

生成された決定木から，複数のキーワードが出現した場合に評判が決定される例に注目することで，評判の原因を分析する．以下にいくつかの例を示す．

Rule 1:A[*2], 画面 → 評判 (良)
Rule 2:壊れる，ハードディスク，B[*3]→ 評判 (悪)
Rule 3:デザイン，キーボード → 評判 (悪)
Rule 4:軽い，欲しい → 評判 (良)

上記の Rule を満たすコメントを実際のコーパスより調べた．Rule 1 については，ある PC が A という型番にモデルチェンジしたところ，画面が大きくなり非常に見やすくなったという意見が 3 つあった．Rule 2 については，B 社のハードディスクが壊れ，大きな支障をきたしているという不満が 2 つあった．Rule 3 については，デザインはいいがキーボード配列に不満があるという意見が 2 つ，キーボードは打ちやすいがデザインが悪いという意見が 1 つあった．Rule 4 については，軽いノート PC が欲しかったので満足しているという意見が 2 つ，もっと軽い PC が欲しかったという不満が 1 つあり，Rule 4 では、評判が逆となってしまう例もあった．

このように，決定木から得られたルールを分析することで，学習に利用したコーパス中にどのような意見が含まれているのかを調べる手がかりとすることができ，掲示板のコメントを総当りで読むよりも評判を分析する効率の向上に繋がる．

∗2　PC のブランドおよび型番名
∗3　会社名

4.2 入力の長さの違いによる決定木の変化

家電や PC に関する掲示板では,「〜はいい．しかし，〜なところが気にくわない.」といった様に，1 つのコメントで評判に関しては正反対のことを述べていることがある．したがって，評判情報を抽出する際には C4.5 に与える入力を 1 コメントごとではなく，1 文ごとに入力を与えたほうが良い可能性がある．[藤村 03] では，1 文を入力単位として同様の実験を行った．

生成された決定木を比較すると，C4.5 への入力単位が 1 文の場合ではあるキーワードが現れたら，評判が決定するといったことが繰り返され，決定木が一方向にのみ成長してしまった．一方，入力単位を 1 コメントとした場合では，同一の深さに多数のノードが存在する複雑な決定木が生成された．これは，1 文では 1 つの入力につき持っている属性値，つまり，キーワードは高々1 つか 2 つ程度であり，属性が疎なために持っている属性の重みが非常に大きくなってしまっていたからだと考えられる．

したがって，評判の分析を行う際には一つのキーワードだけで評判が決定するようでは，原因の分析を行うことが難しいので，決定木が複雑になるよう 1 コメントを入力の単位としたほうが良い．

5. 今後の課題

本手法を評判情報を抽出するという部分のみで考えると，C4.5 より SVM を用いたほうが精度・再現率が高くなると考えられる．従って，SVM の有効性を調べる必要がある．

一方，意思決定支援への応用を考えると，C4.5 の精度・再現率のさらなる向上が必要となる．掲示板の文章では，表記のゆれが大きいという問題があり，これが精度・再現率を下げる原因となっているので何らかの補正が必要であると考えられる．また，[平 00] により C4.5 では，属性数を増やしすぎると過学習によって精度が落ちることが示されているので，特に $tf \cdot idf$ による重要語の最適な属性数を調べる必要がある．

6. まとめ

本稿では，機械学習 (C4.5) を用いて，評判を抽出するルールを発見し，決定木を分析することによって意思決定支援へ応用する手法を提案した．さらに，評価実験を行い意思決定支援への応用の一例を述べた．

参考文献

[Quinlan 95] J .R .Quinlan 著，古川康一　監訳:AI によるデータ解析，TOPPAN PUBLISHING(1995).

[立石 01] 立石健二, 石黒義英, 福島俊一:インターネットからの評判情報検索, 情報処理学会研究報告,NL-144-11,pp.75-82(2001)

[藤村 03] 藤村滋, 松村真宏, 石塚満:機械学習による電子掲示板からの評判情報抽出, 情報処理学会第 65 回全国大会講演論文集,pp[4-451]-[4-452](2003)

[平 00] 平博順, 春野雅彦:Support Vector Machine によるテキスト分類における属性選択, 情報処理学会論文誌,Vol.41,No.4,pp.1113-1123,(2000)