Semi-Supervised な学習手法による評価表現分類

鈴木 泰裕[†]　高村 大也[††]　奥村 学[††]
[†] 東京工業大学 大学院総合理工学研究科　[††] 東京工業大学 精密工学研究所
yasu@lr.pi.titech.ac.jp　{takamura,oku}@pi.titech.ac.jp

1 はじめに

近年, MovableType などの Weblog 作成支援ツールや, ココログなどの Weblog ホスティングサービスの普及により, Web 上で Weblog を公開する人が急増している. そのため, 個人が発信する有用な情報源の一つとして Web 上の文書に注目が集まっている. 個人が発信する情報の中でも現在特に重要視されている情報の一つは評判情報である. Web 上に散在している評判情報を一括して扱えるようにすることは, 一般の人々にとっての意思決定や, 企業にとっての危機管理などの支援になる. Weblog に含まれる評判情報を扱うには, テキスト中の評価表現を抽出し, その表現が肯定・否定のどちらの感情を含んでいるかを特定することが不可欠である. 本研究で対象にしている評価表現とは, 例えば「この HDD は容量が大きい.」という評判情報を含む文の中の, どの対象 (HDD) についての, どの部分 (容量) が, どうなのか (安い) という評判情報の核となる三要素からなるものである.「大きい」という語だけでは肯定・否定を判定することはできないが, この三つ組単位で扱うことにより判定が可能になる. 評価表現の極性判定を行うために, 従来手法の多くでは, 人手により肯定・否定の分類を行い, 評価表現辞書を作成していた. しかし, 評価表現によっては評価対象ごとに込められている肯定・否定の感情が違ってくる場合があり, 評価対象などの情報を含めた辞書を作ろうとすると, 莫大なコストがかかる. そこで本研究では, Web 上から評判情報を収集・分析するための技術の一つとして, 評判情報の核となる評価表現を抽出しつつ, 肯定的な表現であるか, 否定的な表現であるかを判定する semi-supervised な手法を提案する.

2 提案手法

我々は, 評価表現はある種の周辺情報を伴って現れるという仮定を置いている. そして, 評価表現の分類結果により新たな評価表現や周辺情報を発見することができ, また獲得された評価表現や周辺情報により未知の評判情報を正しく分類することができると考えている.

このようなブートストラッピング的な発想を実現するために, 教師付き学習手法を EM アルゴリズムで補強する semi-supervised な方法を適用する. ここでは, 教師付き学習手法として, ナイーブベイズ分類器を用いる. ナイーブベイズ分類器を用いたのは, EM アルゴリズムと組合せることにより, 文書分類で高い性能を発揮することが Nigam らにより示されているからである [1]. 以下, 本研究で用いた semi-supervised な方法を説明する.

2.1 ナイーブベイズ分類器による評判情報分類

ここではまずナイーブベイズ分類器 (Naive Bayes classifiers) の一種である多項モデル [2] について説明する.

多項モデルでは, カテゴリ c が与えられたときに, 事例 $\mathbf{x}$ が生起する確率は,

$$P(\mathbf{x}|c,\theta) = P(|\mathbf{x}|)|\mathbf{x}|!\prod_w \frac{P(w|c)^{N(w,\mathbf{x})}}{N(w,\mathbf{x})!} \quad (1)$$

となる. ここで, $P(|\mathbf{x}|)$ は長さ $|\mathbf{x}|$ の文が生起する確率であり, $N(w,\mathbf{x})$ は文 $\mathbf{x}$ 中での素性 w の出現頻度である. 文の生起はイベントのセットとしてモデル化され, そのイベントでは, 単語は全語彙の中から選び出される.

ナイーブベイズ分類器を評判情報分類に適用した場合, 各発言が事例 $\mathbf{x}$ に相当し, カテゴリ c は, 肯定的評価, 否定的評価, 非評価のいずれかの値をとる. 使用される素性は, 評価表現やその周辺に出現する単語などであり, 第 3 章で詳述する.

2.2 ナイーブベイズ分類器と EM アルゴリズムの組合せ

EM アルゴリズムはいくつかの変数 (隠れ変数と呼ばれている) が観測できない状況で, モデルを最尤推定する手法である [3]. Nigam ら [1] はナイーブベイズ分類器と EM アルゴリズムを組み合わせることを提案している.

ナイーブベイズ・モデルの式において, 関係ない要素を無視すると, 次の式を得る：

$$P(\mathbf{x}|c,\theta) \propto \prod_w P(w|c)^{N(w,\mathbf{x})}, \quad (2)$$

$$P(\mathbf{x}|\theta) \propto \sum_c P(c)\prod_w P(w|c)^{N(w,\mathbf{x})}. \quad (3)$$

以降, モデルのパラメータ群をまとめて θ と表す.

c を隠れ変数とし, ディリクレ分布をパラメータの事前分布とすると, 対数尤度の隠れ変数に関する期待値 (Q 関数) は次のように定義できる：

$$Q(\theta|\bar{\theta}) = \log(P(\theta)) + \sum_{\mathbf{x}\in D}\sum_c P(c|\mathbf{x},\bar{\theta}) \times \log\left(P(c)\prod_w P(w|c)^{N(w,\mathbf{x})}\right). \quad (4)$$

ここで，$P(\theta) \propto \prod_c (P(c)^{\alpha-1} \prod_w (P(w|c)^{\alpha-1}))$ であり，また，α はハイパーパラメータ，さらに D はモデルの推定に用いられる事例の集合である．

この Q 関数より，次の EM 計算式が得られる：
E-ステップ:

$$P(c|\mathbf{x}, \bar{\theta}) = \frac{P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})}{\sum_c P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})}, \quad (5)$$

M-ステップ:

$$P(c) = \frac{(\alpha-1) + \sum_{\mathbf{x}\in D} P(c|\mathbf{x}, \bar{\theta})}{(\alpha-1)|C| + |D|}, \quad (6)$$

$$P(w|c) = \frac{(\alpha-1) + \sum_{\mathbf{x}\in D} P(c|\mathbf{x}, \bar{\theta})N(w, \mathbf{x})}{(\alpha-1)|W| + \sum_w \sum_{\mathbf{x}\in D} P(c|\mathbf{x}\ , \bar{\theta})N(w, \mathbf{x})}. \quad (7)$$

ここで，$|C|$ はカテゴリ数，$|W|$ は単語数を表す．ラベル付き事例については，式 (5) は使用されない．その代わりに，c が事例 $\mathbf{x}$ のカテゴリならば $P(c|\mathbf{x}, \bar{\theta})$ は 1.0 とし，そうでなければ 0 とする．

統計物理学的視点に基づいた EM アルゴリズムの変種に，Deterministic Annealing EM (DAEM) がある [4, 5]．この変種では，モデルの複雑さを調整することができる．DAEM は，E-ステップで式（5）の代わりに次式を使用することで実現できる：

$$P(c|\mathbf{x}, \bar{\theta}) = \frac{\{P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})\}^\beta}{\sum_c \{P(c|\bar{\theta})P(\mathbf{x}|c, \bar{\theta})\}^\beta}. \quad (8)$$

ここで，β はモデルの複雑さを決めるハイパーパラメータで，値が大きいほどモデルは複雑になる．ラベルなしデータに対してラベルありデータが極端に少ないと，学習を繰り返していくうちにラベルなしデータの影響が強くなりすぎて，結果が悪くなってしまうことがある．そのため，$\lambda(0 \leq \lambda \leq 1)$ を用いて，ラベルなしデータの影響が小さくなるように式（4）の右辺の第 2 項を次式と入れ換える：

$$\sum_{\mathbf{x}\in D^l} \sum_c P(c|\mathbf{x}, \bar{\theta}) \log\left(P(c) \prod_w P(w|c)^{N(w,\mathbf{x})}\right)$$
$$+\lambda \sum_{\mathbf{x}\in D^u} \sum_c P(c|\mathbf{x}, \bar{\theta}) \log\left(P(c) \prod_w P(w|c)^{N(w,\mathbf{x})}\right).$$

ここで，D^l はラベル付きデータ，D^u はラベルなしデータである．この式が示すように，λ の値が小さいほど，ラベルなしデータの影響が小さくなる．

この新たな Q 関数を用いて導出したアルゴリズムを使用した．Q 関数の値の変化が十分に小さくなることを終了条件とした．

2.3 Support Vector Machines と フィッシャーカーネル

フィッシャースコアは確率モデルの対数尤度をパラメータで微分してできたベクトルである．semi-supervised な方法で確率モデルを推定し，これを SVM(Support Vector Machines) 等の高性能な分類器と組み合わせることにより，性能を向上させることができるという報告がある [6]．今回は評価表現を分類するタスクにおいて，これと同様な方法を利用した．

2.4 ハイパーパラメータの推定

EM 学習を行うにあたって，λ と β というハイパーパラメータの値を設定する必要があるが，分類性能はこのハイパーパラメータによって大きく依存する．最適なハイパーパラメータを推定するための方法としては，あらゆる組み合わせについて，交差検定による評価実験を行い，最も結果が良かったときの組み合わせを採用するという決定の仕方が考えられる．しかし，ハイパーパラメータの多数の組み合わせについて毎回学習とテストを行うのは，計算コストがとても高いため，テスト事例を含めたデータで繰り返し学習を行い，最後にテスト事例の影響を除いて生成したモデルでテスト事例を分類し，正解率を測定するということを行い，その正解率が最も高かった組み合わせを最適値と予測することにした．

3 評価実験・結果と考察

3.1 評価実験のためのデータの準備

評価実験に使用するデータの準備として，奥村ら [7] が開発したシステム (blogWatcher) により収集した blog 記事 (html の一部) から，文を切り出し，係り受け解析を行い，三つ組とその周辺情報を抽出した．収集した記事はヒューリスティックスにより文単位に分割し，係り受け解析を行う．係り受け解析には，CaboCha[1]を用いた．次に，評価表現の候補すなわち評価対象・属性・評価語の三つ組の候補を収集した．評価語の候補としては，形容詞・形容動詞・動詞の「ある」を考える．そして，この評価語の候補に係る文節の中から，評価対象，属性を見つけ，対象・属性・評価語の三つ組を抽出した．ランダムにサンプリングした約 200 事例に関して調査した結果，対象・属性の誤抽出があった事例が 22.0%，形態素解析・構文解析誤りがあった事例が 14.0%あったが，対象・属性が誤抽出されている事例も，評価語候補に注目することで肯定的/否定的/非評価の判定はできるため，後述の実験では正しいデータと同等に扱った．

3.2 考慮する周辺情報

機械学習で素性として用いる周辺情報として，手がかりになり得ると考えられる以下の情報を用いた．

1. 文に含まれる感動詞
2. 文に含まれる丸括弧の中の形態素
3. 三つ組いずれかに係る文節内の形態素
4. 評価語候補が係る文節内の形態素
5. 評価語候補と同じ文節に含まれる形態素
6. 文の直前にある顔文字のカテゴリ
7. 文の直後にある顔文字のカテゴリ

[1]公式 HP:http://chasen.org/~taku/software/cabocha/

表 1: 各手法の正解率

手法	正解率 (%)
Baseline	47.5
NaiveBayes	76.0
SVM	76.6
NaiveBayes+EM	77.1
SVM+NaiveBayes+EM	77.6

以前行った調査 [8] により特に有用であると判断した周辺情報について特別扱いし，他の周辺情報との 2-gram を素性とした．具体的には「〜だが」といった逆接表現や「〜すぎる」「〜ない」といった表現である．顔文字のカテゴリは，田中ら [9] の開発した顔文字抽出・分類手法を利用し，文中から顔文字を抽出し，喜んでいる，悲しんでいる，怒っている，驚いている，動作を表している，苦笑しているの 6 通りに分類し，そのカテゴリ名を素性とした．これらの素性に評価表現自体を加えたものを学習時の素性とした．

3.3 評価実験用データ

Web 上の文書には非文が多く，口語的表現も多く, 正確な係り受け解析が困難であるので, 得られたデータにノイズが多く含まれるという問題がある．また周辺情報が少ない場合には分類が困難であると考えられる．そのため，そのような事例をフィルタリングする規則を適用した．更に，出現回数が少ない素性は効果が無く，ノイズになる可能性が高いため，2 回以上出現した素性のみを使用することにした．フィルタリングの結果，blog 約 50 万記事 (1 記事は 1 日分) から抽出した約 260 万の事例は，35765 に減少した．人手によりランダムサンプリングした事例に対してラベル付けを行った．その結果，1061 事例に対してラベルが付与された．ラベルの内訳は非評価が 69(6.5%)，肯定的表現が 504(47.5%)，否定的評価が 488(46.0%) であった．このラベルが付与された 1061 事例のデータを用いて提案手法の評価実験を行った．ナイーブベイズ分類器のハイパーパラメータ α については 2.0 に固定した．EM アルゴリズムで取り込まれるラベルなしデータは，34704 事例である．ラベル付きデータ中で最も該当する事例が多いのは肯定的評価のクラスであるため，全ての事例について肯定的評価のクラスであると予測したときの正解率 47.5% が最も粗いベースラインと考えられる．

3.4 実験

各手法において 10 分割交差検定により評価実験を行った結果を表 1 に示す．ハイパーパラメータは前述の手法により推定した値を用いている．ナイーブベイズと SVM の両手法とも，ラベルなしデータを学習に取り入れることで，分類正解率が向上しており，ラベルなしデータの利用が有効であることがわかる．また，λ が 0.01，β が 0.9 という最適な組に固定して各分割で実験を行った場合のナイーブベイズ+EM アルゴリズムの正解率は 77.4%であり，最悪の組に固定した場合の正解率は 50.9%であったため，本手法により推定したハイパーパラメータの値による正解率 (77.1%) は最適な組の場合にかなり近く，推定はうまく行えていると言える．

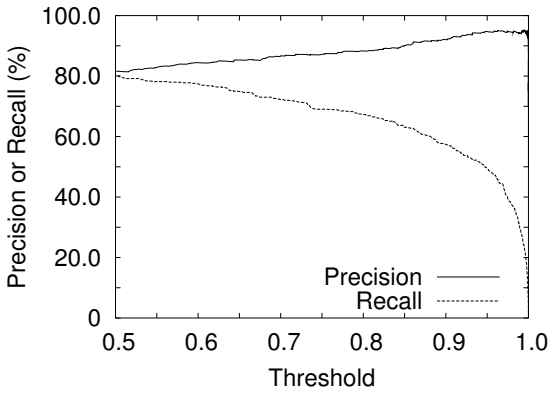

図 1: 確信度の下限に対する肯定的クラスの Precision と Recall

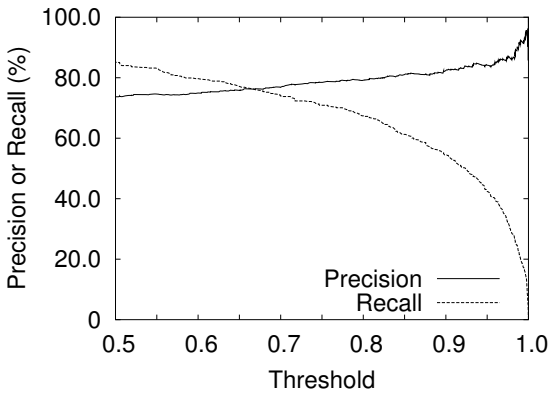

図 2: 確信度の下限に対する否定的クラスの Precision と Recall

2.2 節における式 (8) の最終的な値は，各事例の非評価・肯定的評価・否定的評価クラスへの属し易さになっていると考えることができる．そのため，ある事例 $\mathbf{x}_i$ がクラス c_i に分類されたときの $P(c_i|\mathbf{x}_i)$ の値を確信度と呼ぶことにする．肯定的な評価であると分類された事例と否定的な評価であると分類された事例に関して，採用する事例の確信度の下限に対する，Precision(精度) と Recall(再現率) をグラフにしたのが図 1 と図 2 である．肯定的評価のクラスでは確信度が 0.85 以上の事例に限定すれば 90%という高い Precision を得られている．否定的評価のクラスでは 90%の Precision を得ようとすれば，確信度が 0.98 以上の事例に限定する必要があり，否定的評価クラスへの分類の方が難しいということがわかる．

学習により各素性について，各クラスの手がかりへのなり易さが求められるが，喜んでいる顔文字や「(爆)」が肯定的表現の手がかりになり易く，怒っている顔文字・悲しんでいる顔文字・苦笑している顔文字や「(泣)」が否定的表現の手がかりになり易くなっていた．また，「〜なので良い」「〜だからつまらない」というような因果関係の情報も有用な情報となっていた．

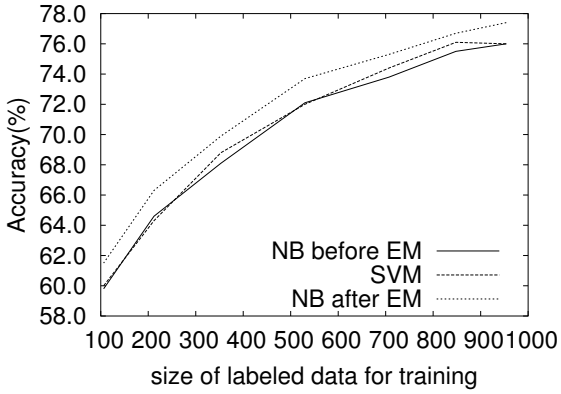

図 3: 訓練データのサイズに対する正解率

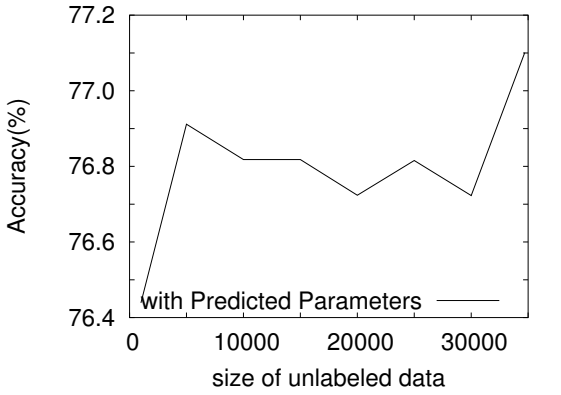

図 4: ラベルなしデータ量に対する分類正解率

誤分類が起きていた事例に関して見てみると，係り受け解析などの基盤技術の精度向上や，学習の際の素性を更に拡張することで，対応できる可能性があると考えられるものであった．

3.5 データサイズの影響

ラベル付き訓練データが更に小さいときに，提案手法がどのような振舞いをするかを観察するため，訓練データサイズを変化させ実験を行った．結果は図 3 である．この図からわかるように，適切なハイパーパラメータが選ばれれば，常に EM アルゴリズムの結果がナイーブベイズ分類器のみの正解率や SVM の正解率を上回っており，ラベル付き訓練データのサイズに関わらず EM アルゴリズムが効果的であることがわかる．

次に，ラベルありデータ量はそのままで，ラベルなしデータ量を変化させた結果を図 4 に示す．ラベルなしデータが 5000 事例だけの場合でも充分な正解率の向上が見られた．途中，ラベルなし事例数が 10000〜30000 の辺りで正解率の伸びが停滞しているが，ラベルなしデータを全て使用した場合が最も正解率が高かった．ラベルなしデータの量を更に増やすことで，性能が向上することが期待できる．ラベルなし事例数が 10000〜30000 の辺りでの結果が安定していない原因に関しては，更に調査を行う必要がある．

4 まとめ

本論文では，Web 上から評判情報を収集・分析するための技術の一つとして，評判情報の核となる評価表現を抽出しつつ，発言全体が肯定的な評価であるか，否定的な評価であるか，あるいは非評価であるかを判定する semi-supervised な手法を提案した．提案手法は，ナイーブベイズ分類器と EM アルゴリズムと SVM の組合せから成っている．実験では，ハイパーパラメータが適切に選択された場合，EM アルゴリズムによりナイーブベイズ分類器の分類正解率が 1.4%向上し，SVM とも組み合わせることにより，分類正解率は更に 0.5%向上した．また，訓練データサイズを変化させても，EM アルゴリズムによる semi-supervised な学習手法は効果的であった．そして，EM アルゴリズムを適用する際のハイパーパラメータの最適な組み合わせを自動的に決定する手法も有効に働くことがわかった．

今後の課題としては，まず前処理の改善が考えられる．文区切り，対象・属性の抽出の精度向上，照応省略解析の導入などが課題である．それから，分類手法についても改善の余地がある．例えば，ある評価表現が別の評価表現の周辺情報になっている場合，その情報を素性として加え，訓練・分類・素性の更新を繰り返すことで性能を向上させることができると考えている．また，単語のクラスタリングと本手法を組み合わせることも重要な発展方向である．

参考文献

[1] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.

[2] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48, 1998.

[3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, Vol. 39, No. 1, pp. 1–38, 1977.

[4] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. Technical Report AIM-1625, Artifical Intelligence Laboratory, MIT, 1998.

[5] Naonori Ueda and Ryohei Nakano. Deterministic annealing variant of the em algorithm. In *NIPS7*, pp. 545–552, 1995.

[6] Hiroya Takamura and Manabu Okumura. A comparative study on the use of labeled and unlabeled data for large margin classifiers. In *IJCNLP*, pp. 620–625, 2004.

[7] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕. blog ページの自動収集と監視に基づくテキストマイニング. 人工知能学会, セマンティックウェブとオントロジー研究会,SIG-SWO-A401-01, 2004.

[8] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02, 2004.

[9] Yuki Tanaka, Hiroya Takamura, and Manabu Okumura. Extraction and classification of facemarks with kernel methods. In *IUI*, 2005.