EM アルゴリズムを用いた語義判別規則の教師なし学習*

新納浩幸　　○ 高橋篤史

茨城大学 工学部 システム工学科

1 はじめに

本論文では，Nigam らによって提案された EM アルゴリズムを利用した教師なし学習の手法 [3] を語義判別問題に適用する．また教師なし学習によって判別精度を高めるための素性の選択について検討する．

自然言語処理では個々の問題を分類問題として定式化し，帰納学習の手法を利用して，その問題を解決するというアプローチが大きな成功をおさめている．しかしこのアプローチには帰納学習で必要とされる訓練データを用意しなければならないという大きな問題がある．この問題に対して，近年，少量のラベル付き訓練データから得られる分類規則の精度を，大量のラベルなし訓練データによって高めてゆく seed 型の学習が散見される．近年の教師なし学習の多くの手法は，本質的には，判別のための複数観点を利用する．複数観点を利用した代表的な手法として，Co-training[1] と EM アルゴリズムを利用した手法 [3] がある．どちらの手法も本来は文書分類に対して考案されており，語義判別問題に適用できるかどうかは明らかではない．語義判別問題の解決は自然言語処理の中心的な課題であり，これらの手法が語義判別問題に適用できることが望ましい．ここでは EM アルゴリズムの適用可能性について検討する．

EM アルゴリズムを利用した手法は，Naive Bayes と EM アルゴリズムを組み合わせた手法である．本質的には，あるクラス c のもとである素性 f が発生する確率 $P(f|c)$ を求める．Naive Bayes のモデルが使えれば，この確率から分類器を作成できる．ラベル付きデータから $P(f|c)$ は簡単に計算できるが，ラベル付きデータが少量の場合，$P(f|c)$ の信頼性は低い．そこでラベルなしデータを用いて $P(f|c)$ の信頼性を高める．これは全体のデータの発生確率が最大になるように $P(f|c)$ を設定すればよい．この計算に EM アルゴリズムが利用される．この手法も，本質的に，複数観点を利用している．事例は素性のベクトル $(f_1, f_2, \cdots, f_n)$ として表されるので，ある素性 f_i から，その事例のクラス c が判別できたとき，$P(f_j|c)$ の精度を高められるからである．

上記の手法を語義判別問題に適用する場合，判別力の高い素性の組を選択する他に，語義判別に Naive Bayes のモデルが適用できるような素性の組を設定することが鍵と考えられる．本論文では，利用する素性の組が Naive Bayes のモデルの仮定を満たす程度とその素性の組の判別力の 2 点を考慮して，素性を選択する効果について検討する．

*Unsupervised learning of word sense disambiguation rules by using EM algorithm

2 Naive Bayes による語義判別

ある事例 x が素性のベクトルとして，以下のように表現されたとする．

$$x = (f_1, f_2, \cdots, f_n)$$

x の分類先のクラスの集合を $C = \{c_1, c_2, \cdots, c_m\}$ と置く．問題は $P(c|x)$ の分布を推定することである．実際に，x のクラス c_x は以下の式で求まる．

$$c_x = arg \max_{c \in C} P(c|x)$$

ベイズの定理を用いると，

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

なので，結局，以下が成立する．

$$c_x = arg \max_{c \in C} P(c)P(x|c)$$

ここで，$P(c)$ は比較的簡単に推定できる．問題は，$P(x|c)$ の推定だが，これは現実的には難しい．Naive Bayes のモデルは，この推定に以下の仮定を導入する．

$$P(x|c) = \prod_{i=1}^{n} P(f_i|c) \quad (1)$$

$P(f_i|c)$ の推定は比較的容易であるために，結果として $P(x|c)$ が推定できる．Naive Bayes を使った分類がうまくゆくかどうかは，式 1 の仮定をできるだけ満たすような素性を選択することである．文書分類であれば，各素性を各単語の生起に設定することで，Naive Bayes が有効であることが知られている．

語義判別問題でも式 1 の仮定をできるだけ満たすような素性を選択すれば Naive Bayes が利用できる．

本論文では (e1) 直前の単語，(e2) 直後の単語，(e3) 前方の内容語（2 つまで），(e4) 後方の内容語（2 つまで），の 4 つの素性を設定した．例えば，「声」の語義は『意見』という語義と『喉から発声される音』という語義がある．また，「日本国民の声を集めました」という文は以下のように形態素解析される．各行が分割された単語であり，第 1 列が表記，第 2 列が原型，第 3 列が品詞を表す．

日本	日本	名詞-固有名詞-地域-国
国民	国民	名詞-一般
の	の	助詞-連体化
声	声	名詞-一般
を	を	助詞-格助詞-一般
集め	集める	動詞-自立
まし	ます	助動詞
た	た	助動詞

この結果から以下の 4 つの素性ができる．

```
e1=の，e2=を，
e3 ={日本，国民}，e4={ 集める }
```

e3 と e4 の素性は集合になるが，学習の際に以下のように分割する．

e3=日本，　e3=国民，　e4=集める

以上のように設定した素性集合が Naive Bayes が仮定する式 1 をどの程度満たすかを求めることは難しい．ただ，名詞の場合，その単語の右文脈と左文脈はほぼ独立と考えて良い点と，2 単語列を素性として含めない，などを考慮して設定した．

3 EM アルゴリズムによる教師なし学習

分類問題の解決に Naive Bayes が使えれば，Nigam らが提案した教師なし学習が利用できる．そこでは EM アルゴリズムを用いることで，ラベルなしデータを用いて，ラベル付きデータから学習された分類器の精度を向上させる．

ここではポイントとなる式とアルゴリズムだけを示す [3]．

基本となるのは，あるクラス c_j のもとで，素性 f_i が発生する確率 $P(f_i|c_j)$ を求めることである．これは以下の式で求まる．この式は頻度 0 の部分を考慮したスムージングを行っている．

$$P(f_i|c_j) = \frac{1 + \sum_{k=1}^{|D|} N(f_i, d_k) P(c_j|d_k)}{|F| + \sum_{m=1}^{|F|} \sum_{k=1}^{|D|} N(f_m, d_k) P(c_j|d_k)} \quad (2)$$

式 2 の D はラベル付けされたデータとラベル付けされていないデータを合わせた訓練データ全体を示す．D の各要素を d_k で表す．F は素性全体の集合である．F の各要素を f_m で表す．また，$N(f_i, d_k)$ は，訓練事例 d_k に含まれる素性 f_i の個数を表す．ここでの設定では，$N(f_i, d_k)$ は 0 か 1 の値であり，ほとんどの場合 0 である．$P(c_j|d_k)$ は訓練データがクラス c_j を持つ確率である．ラベル付けされたデータに対しては，0 か 1 の値をとる．ラベル付けされていないデータに対して，最初は 0 であるが，EM アルゴリズムの繰り返しによって，徐々に適切な値に更新されてゆく．

式 2 を利用して，以下の分類器が作成できる．

$$P(c_j|d_i) = \frac{P(c_j) \prod_{f_n \in K_{d_i}} P(f_n|c_j)}{\sum_{r=1}^{|C|} P(c_r) \prod_{f_n \in K_{d_i}} P(f_n|c_r)} \quad (3)$$

ここで，C はクラスの集合である．K_{d_i} は訓練事例 d_i に含まれる素性の集合を示す．$P(c_j)$ はクラス c_j の発生確率であり，以下の式で計算できる．

$$P(c_j) = \frac{1 + \sum_{k=1}^{|D|} P(c_j|d_k)}{|C| + |D|}$$

EM アルゴリズムは式 2 を利用して，$P(f_i|c_j)$ を求め (E-step)，次にこの値を利用して式 3 から分類器を作成し，ラベル付けされていない事例 d_i に対して，$P(c_j|d_i)$ を求める (M-step)．この E-step と M-step を交互に繰り返して，$P(f_i|c_j)$ と $P(c_j|d_i)$ を収束するまで更新してゆく．

最終的には式 3 から分類が行える．

4 素性選択の指標値

前述の教師なし学習手法を用いる場合，どのような素性を使うかが重要である．ここでは，判別力の高さ

(P) と Naive Bayes のモデルの仮定を満たす程度 (E) の 2 つの観点が素性選択に影響していると考え，それらを使って，素性選択の指標値を提案する．

まず確率モデルどうしの距離を KL 情報量により測り，E を以下のように定義する．

$$E = \sum_{c} P(c) I(\prod_{i=1}^{n} P(f_i|c), P(x|c))$$

ここで I は真のモデル p と比較対象のモデル q との KL 情報量であり，以下で定義される．

$$I(p;q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

次に判別力の高さ (P) は，ラベル付き訓練データのみから学習できる Naive Bayes の分類器のラベル付き訓練データに対する精度に設定する．

E の値は小さいほどよく，P の値は大きいほどよい．P の値は正規化されているので，E/P を素性選択の指標値とすることにした．

ただし現実的には素性の数が多いと，E の値は計算できない．以後の実験では，素性の数を 2 つに限定して，この指標値の妥当性を示す．

5 実験

単語「声」の語義判別規則の学習に EM アルゴリズムを適用する．「声」の語義は大きく『意見』と『喉から発声される音』という語義がある．ここでは曖昧性をなくすために『意見』という語義とそれ以外の語義という形で 2 つの語義を設定し，前者を a，後者を b とする．

次に毎日新聞の '95 年度 1 年分の記事を形態素解析し，「声」という単語を含む文を取り出した．全部で 7,041 文存在した．次に，そこからランダムに 100 文と 300 文を取り出し，それらの各文の持つ「声」の語義に応じて a または b のラベルを付与した．ラベル付きの 100 文を最初のラベル付き訓練データ L とし，ラベル付きの 300 文を評価のためのテストデータ T とした．残りの 6,641 文がラベルなし訓練データ U である．EM アルゴリズムは 10 回の繰り返しで終了することにした．

素性 e1 ～ e4 を使った場合の結果を図 1 に示す．横軸は EM アルゴリズムの繰り返しの回数，縦軸は学習できた分類器の T に対する正解率 (%) である．最初のラベル付き訓練データだけから得た分類器の精度は 80.7% であったが，EM アルゴリズムを用いることで，84.0% まで向上した．

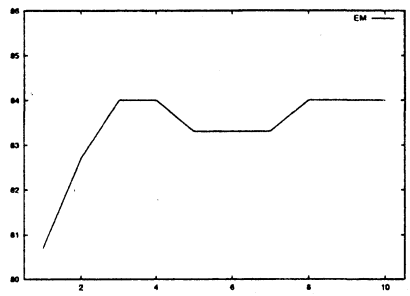

図 1: EM アルゴリズムによる教師なし学習

次に素性 e1 ～ e4 のうちから 2 つの素性だけを使った学習を行ってみる．まず利用する 2 つの素性が Naive Bayes の仮定を満たす程度 (E) を測る．このために L から得たデータを使い，確率モデルを推定した．現れない事例は頻度 0.1 にとった．また利用する 2 つの素性の組の判別力 (P) を測った．これは L から Naive Bayes を用いた分類器を作成し，L における正解率である．以上より提案した素性選択の指標値 E/P の値を算出した．次に各々の素性の組を用い，L のみから Naive Bayes による分類器の T に対する精度と，EM アルゴリズムを利用した教師なし学習を行って最終的に得られた分類器の T に対する精度を求めた．結果を表 1 にまとめる．表から E/P の値が最も小さい e3 と e4 の組が，最も効果的に学習できていることがわかる，

6 考察

本実験により，語義判別問題に対しても，EM アルゴリズムを使った教師なし学習が適用可能であることが示された．ある種の単語に対しては，この手法により精度向上が図れる．現在，我々は SENSEVAL2 の日本語翻訳タスク [4] で課題とされた単語に対して EM アルゴリズムを使った教師なし学習を試みているが，そこでも約半数は精度が向上された．しかし精度が逆に悪くなるケースも存在する．本実験でも e1 と e2 だけで学習させると，精度は悪くなる．しかも，この 2 つの素性の組は Naive Bayes の仮定を満たす程度 (E) が最も良い．つまり単純に E の値だけでは利用すべき素性を判断できない．また単純に想定した素性をすべて使えば正解率が向上するわけでもない．そのために

素性	E	P	E/P	Naive Bayes	Naive Bayes + EM
e1 + e2	0.412	0.85	0.485	79.3 %	77.3 %
e1 + e3	0.671	0.98	0.685	82.3 %	85.3 %
e1 + e4	0.593	0.92	0.645	84.3 %	85.3 %
e2 + e3	0.628	0.96	0.654	76.7 %	77.0 %
e2 + e4	0.673	0.91	0.740	78.0 %	75.7 %
e3 + e4	0.460	0.99	0.464	78.7 %	86.3 %

表 1: 素性選択の指標値と正解率

頑健性の高い素性の選択法は重要な課題である.

本実験では 2 つの素性に限定して, Naive Bayes の仮定を満たす程度 (E) と, 素性の組の判別力 (P) から素性選択の指標値 (E/P) を与えた. ただし本論文で用いた教師なし学習の手法は, 利用する素性が 3 つ以上でも全く構わない. その場合, どのようにして E を測ればよいかは今後の課題である. また E と P が素性選択に関わっていると思われるが, 指標値を E/P の形で与えることが適切かどうかも今後検討する必要がある. 例えば, 素性 e1, e3, e4 の 3 つを使った教師なし学習の結果を表 2 に示す. これが今回行った実験の中でも最もよい値を出した. Naive Bayes の仮定を満たす程度から考えれば, 素性 e1, e3, e4 の 3 つ組はよい選択とは思えないので, E/P では適切な指標値を与えられていない.

素性	Naive Bayes	Naive Bayes + EM
e1 + e2 + e3 + e4	80.7 %	84.0 %
e3 + e4	78.7 %	86.3 %
e1 + e3 + e4	83.3 %	87.0 %

表 2: 精度の高い素性の組

複数観点を利用した手法としては, 本論文で用いた EM アルゴリズムを利用した手法の他に Co-training [1] がある. この 2 つの手法を比べた場合, Co-training は独立な 2 つの素性集合を設定しなければならないという問題がある. 一方, EM アルゴリズムを利用した手法は, データの発生源や素性に課しているモデルが厳しい. このため Co-training の方が応用範囲が広く現実的な手法と言える. また文書分類に限れば, Co-training の方が EM アルゴリズムを利用した手法よりも優れていたという報告もある [2]. しかし語義判別のような多値分類問題を扱う場合, Co-training は独立な 2 つの素性集合という問題以外に, 素性の一貫性という条件が必要であり, その適用は難しい. 一方, EM アルゴリズムを利用した手法は, 原理的には分類問題が多値であっても影響はない. このため多値の分類問題への適用の観点から見れば, EM アルゴリズムを利用した手法の方が現実的である.

7 おわりに

本論文では, Nigam らによって提案された EM アルゴリズムを利用した教師なし学習の手法を語義判別問題に適用した.「声」の語義判別の実験では, この手法を用いることで, 判別の精度を向上させることができた. また Naive Bayes の仮定を満たす程度 (E) と, 素性の組の判別力 (P) から素性選択の指標値を提案した. 2 つの素性を使う場合には, この指標値がうまく機能していることが確認できた. 今後はより頑健性の高い素性の選択法を検討し, 多値の語義判別問題へ適用していきたい.

参考文献

[1] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *11th Annual Conference on Computational Learning Theory (COLT-98)*, pp. 92–100, 1998.

[2] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-trainig. In *9th International Conference on Information and Knowledge Management*, pp. 86–93, 2000.

[3] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. In *Machine Learning*, Vol. 39, pp. 103–134, 2000.

[4] 黒橋禎夫, 白井清昭. SENSEVAL-2 日本語タスク. 電子情報通信学会, 言語とコミュニケーション, NLC2001-36 ~48, pp. 1–8, 2001.