技術早分かり

最大エントロピー法と自然言語処理

茨城大学 工学部 システム工学科　新納浩幸
E-mail : shinnou@dse.ibaraki.ac.jp

1　はじめに

ある事例がどのクラスに属するかを決定する分類問題は、人工知能や統計学の分野で活発に研究されている。自然言語処理の分野では、個々の問題を分類問題に変換することで、それらの研究を応用する。

分類問題に対する手法は種々あるが、自然言語処理の分野では、近年、最大エントロピー法 [10] を利用した研究が散見される。最大エントロピー法は従来主流であった決定木よりも優位な結果を出す傾向があり、今後ますます利用されるであろう。

しかし最大エントロピー法の優位性にも関わらず、この手法を利用している研究者はごく一部である。おそらく最大エントロピー法が一見難解であり、取っつきが悪いからであろう。本稿では最大エントロピー法を概説する。特に応用することを念頭に置き、実装の手引きとなることを目的とする。そのため理論面の説明は省いている。理論面での解説は文献 [17] が平易で詳しい。

2　分類問題に対する最大エントロピー法

自然言語処理の様々な問題は分類問題に変換できる。そして、最大エントロピー法は、分類問題に対する帰納学習手法の１つととらえることができる。

入力データ d がクラスの集合 $\{C_1, C_2, \cdots, C_n\}$ のうちどのクラスに属するかを決めるのが分類問題である。この問題は非常に汎用的な形式であり、あらためて考えてみると世の中の様々な問題はこのタイプの問題と見なすことができる。そして、当然、自然言語処理の様々な問題も、この問題に変換できる。例えば、ある文中の bank の語義が「銀行」か「土手」かを決める語義選択の問題は、その bank が現れた文脈を入力データとし、クラスの集合を { 銀行, 土手 } とした分類問題と見なすことができる。分類問題は、実質、どのクラスに属するかを決めるための規則を作る問題である。この規則を人間が手作業で作っても良いが、それはコストの面も含めて困難なケースが多く、機械的に作成できればすばらしい。そのための研究が機械学習である。

分類問題は、通常、帰納学習の手法によって解決できる。帰納学習とは、事例とその正解に当るクラスの組を多数集めた訓練データを用意し、その訓練データからクラス判別の規則を機械的に得る手法である。上記の例で言えば、 bank の現れた文を多数用意し、その文中の bank の語義が、「銀行」か、「土手」かの正解を付与したものが訓練データとなる。帰納学習手法にも様々な手法があるが、どの手法がよいかは、問題に依存するため一概には言えない。ただし、自然言語処理では決定リスト [13] や決定木 [8] が比較的よく用いられてきた。そして近年、本稿で解説する最大エントロピー法が注目されている。その理由の１つは、最大エントロピー法がデータスパースネスの問

題に強い手法であるからである。自然言語処理で行われるコーパスからの知識獲得では、データスパースネスが深刻な問題であるため、最大エントロピー法は自然言語処理の様々な問題を解く際に、他の学習手法に比べて、優位な結果を出す傾向がある。

3 最大エントロピー法の定式化

観測される文脈の集合を B 、クラスの集合を A としたとき、自然言語処理の分類問題は、$cl : B \rightarrow A$ なる関数 cl を作成する問題である。最大エントロピー法の cl に対応する規則の表現形式は、文脈を条件としたクラスに対する条件付き確率の形をとる。すなわち、$b \in B$ 、$a \in A$ に対する $P(a|b)$ が最大エントロピー法が与えるクラス判別規則である。

最大エントロピー法を実装するには、文脈述語関数とそれに対応する素性関数が必要となる。文脈述語関数 cp とは、$cp : B \rightarrow \{true, false\}$ なる関数である。この文脈述語関数を使って素性関数 f が以下のように定義される。

$$f_{cp,a'}(a,b) = \begin{cases} 1 & if \quad a = a' \ , \ cp(b) = true \\ 0 & otherwise \end{cases}$$

具体的に素性関数は、$cp(b) = true$ であるときに、$a = a'$ であるという特徴を表している。この特徴がクラス判別を行うときの鍵になっており、その特徴をいかにうまく設定するかが最大エントロピー法の成否を握っている。

訓練データは $T = \{(\omega_1, O(\omega_1)), (\omega_2, O(\omega_2)), \cdots, (\omega_N, O(\omega_N))\}$ と表せる。ここで ω_i は事例 (a_i, b_i) を表し、$O(\omega_i)$ は事例 (a_i, b_i) の頻度を表す。また用意した素性関数の集合（素性集合と呼ぶ）を $F = \{f_1, f_2, \cdots, f_K\}$ とする。この2つの集合 T と F から $P(a|b)$ を構築するのが、最大エントロピー法である。結論のみ述べると、$P(a|b)$ は以下の式で求まる。

$$P(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{K} \alpha_j^{f_j(a,b)} \tag{1}$$

$$Z(b) = \sum_{a} \prod_{j=1}^{K} \alpha_j^{f_j(a,b)}$$

式1の α_j は素性パラメータと呼ばれる実数値であり、素性関数 f_j に対する重みを表している。素性関数 f_j はクラスを判別すための特徴と見なせたが、素性パラメータはその特徴がどれくらい強いかを表している。大きい値ほど、その特徴がクラス判別に有効であることを示している。式1で未知の数はこの素性パラメータである。つまり最大エントロピー法を利用するには、素性集合を設定し、各素性関数に対する素性パラメータを求めれば良いのである。

式1がどのように導出されるのかは少し複雑である。最大エントロピー法を利用したいだけであれば理解する必要はない。ただしその場合、何が最大のエントロピーなのかという、最大エントロピー法の名前の由来が分からない。この点だけ概略述べておく。与えられた制約を満たすモデルの中で、もっとも一様な分布を選ぼうというのが最大エントロピー法のアイデアである。この一様な分布の測定方法としてモデルのエントロピーを使う。もっとも一様な分布というのは、このエントロピーが最大のものである。

4 素性パラメータの推定

最大エントロピー法の処理の核は、素性パラメータの推定である。これは一般に以下の GIS (Generalized Iterative Scaling) アルゴリズム [6] と呼ばれる反復スケーリング法で行える。

【GIS アルゴリズム】

step 1 補完定数 C、補完素性 f_{K+1} を設定する。

step 2 すべての素性パラメタの初期値 $\alpha_i^{(0)}$ $(i = 1 \sim K+1)$ を 1 にする。

step 3 α_i を以下の式で更新する。

$$\alpha_i^{(n+1)} = \alpha_i^{(n)} + \frac{1}{C} log \frac{E_{\hat{P}}[f_i]}{E_P[f_i]}$$

step 4 step 3 を収束するまで繰り返す。

step 1 の補完定数 C、補完素性 f_{K+1} は以下で定義される。

$$C = \max_{a,b} \sum_{i=1}^{K} f_i(a,b)$$

$$f_{n+1}(a,b) = C - \sum_{i=1}^{K} f_i(a,b)$$

次に step 3 の $E_{\hat{P}}[f_i]$ であるが、これは確率 $\hat{P}$ に対する素性の期待値を表し、以下の式で計算できる。

$$E_{\hat{P}}[f_i] = \sum_{a,b} \hat{P}(a,b) f_i(a,b)$$

この式の中の $\hat{P}(a,b)$ は訓練データ中でのデータ b とクラス a の同時確率であり、以下の式で計算できる。

$$\hat{P}(a,b) = \frac{O(a,b)}{\sum_{a,b} O(a,b)}$$

最後に $E_P[f_i]$ であるが、これは確率 P に対する素性の期待値を表し、以下の式で計算できる。

$$E_P[f_i] = \sum_{a,b} P(a,b) f_i(a,b)$$

この式の中の $P(a,b)$ はデータ b とクラス a の同時確率であり、決定することはできない。これを以下の式で近似する。

$$P(a,b) = \hat{P}(b) P(a|b) \tag{2}$$

式 2 の中の $\hat{P}(b)$ は前述した $\hat{P}(a,b)$ を利用して、以下の式で計算できる。

$$\hat{P}(b) = \sum_{a} \hat{P}(a,b)$$

また式 2 の中の $P(a|b)$ は最大エントロピー法で求めようとしている確率分布である。つまり、GIS アルゴリズムの繰り返しの中でその時点での α_i を用いて、式 1 により設定された確率分布である。

一見単純に見える GIS アルゴリズムであるが、実際には膨大な計算量が必要とされる。$E_P[f_i]$ は定数なので、1 回計算すればすむが、問題は $E_P[f_i]$ である。この式の計算量は $O(|A||B|)$ である。これを各 i について計算するので、GIS アルゴリズムの 1 回の反復の計算量は $O(|A||B||F|)$ となり、この計算量は膨大である。このために以下の 2 つの点に注意して計算量を小さくする工夫がとられる [15]。

- 文脈データ b について $cp_i(b) = true$ となるような i の集合を F_b とする。文脈データ b_1 と b_2 において、F_{b_1} と F_{b_2} が等しいなら、$P(a|b_1)$ と $P(a|b_2)$ は同一となる。
- 訓練データ (a_1, b) と (a_2, b) において、これらに対して 1 を返す素性関数の集合が同一ならば、$P(a_1|b)$ と $P(a_2|b)$ は同一となる。

素性パラメータを推定するには、GIS アルゴリズムの他に IIS (Improved Iterative Scaling) アルゴリズム [7] という反復スケーリング法も利用される。IIS アルゴリズムは GIS アルゴリズムで必要な補完定数や補完素性を導入せずにすむ。IIS アルゴリズムの特殊ケースが GIS アルゴリズムという関係になっている。IIS アルゴリズムでは各反復で求まるパラメータの増分が、ある方程式の解になっており、その解をニュートン法などの数値解析的な手法により求めることで行われる。

実装上の注意として、反復の回数と訓練データの間引きについて述べておく。まず反復の回数であるが、GIS や IIS の反復スケーリング法は収束が遅い場合があり、収束の条件を増分の値で設定するとなかなか止まらない場合がある。現実的には 100 から 200 回の反復回数で反復を終了させる。次に訓練データの間引きであるが、訓練データ T をそのまま利用すると、頻度の低いデータが悪影響を及ぼす。そのため、頻度の低いデータは予め訓練データから取り除いておく間引き処理が行われる。どの程度の頻度以下のものを間引けばよいかについては明らかではない。頻度 10 程度を使うことが多い。

5 素性関数の選択

最大エントロピー法によって推定される確率分布の品質は利用する素性集合に大きく依存する。したがって、利用する素性集合をどのよう決めるかが最大エントロピー法の最も大きな問題である。

素性集合を決定するアルゴリズムとして素性選択アルゴリズム [1] が知られている。これは以下の手順で行われる。

【素性選択アルゴリズム】

step 1 素性関数の候補の集合 S を作成する。

step 2 求めるべき素性集合を $F = \phi$ とおく。

step 3 $\Delta L(f_i)$ を最大にするような S 中の要素 f_i を求め、それを F に追加する。

step 4 最大の $\Delta L(f_i)$ がある閾値以下になるまで、step 3 を繰り返す。

ここで $\Delta L(f_i)$ だが、これは確率モデルの対数尤度の変化量を表す。確率モデル P の対数尤度 $L(P)$ とは以下の式で表せる。

$$L(P) = \sum_{t,h} \hat{P}(t,h) log P(t|h)$$

つまり、素性集合が F である場合に、F を用いて最大エントロピー法から計算できる確率モデル P の対数尤度 $L(P)$ と、素性集合が $F \cup \{f_i\}$ である場合に、$F \cup \{f_i\}$ を用いて最大エントロピー法から計算できる確率モデル P_i の対数尤度 $L(P_i)$ との差が $\Delta L(f_i)$ である。

step 3 では、$F \cup \{f_i\}$ に対する確率モデルを求めるために、S の要素数の回数だけ GIS アルゴリズムを用いなくてはならず、膨大な計算が必要である。このため、様々な効率化手法が提案されている。一般的に行われる効率化は、上記のアルゴリズムの step 3 において、すでに存在する素性集合 F の素性パラメータの集合は、素性関数を 1 つ加えても、変化しないと仮定することである。このように仮定すれば、step 3 において、f_i を F に加えた場合に、GIS アルゴリズムで求める素性パラメータは f_i に対する素性パラメータ α_i だけで済む。

その他の効率化の工夫として、白井は素性の独立性という概念を定義している [15]。素性 f_1、f_2 が独立とは f_1 の値が 1 になるようなクラスの集合と、f_2 の値が 1 になるようなクラスの集合とに交わりがなく、しかも f_1 の値が 1 になるようなデータの集合と、f_2 の値が 1 になるようなデータの集合とにも交わりがないことである。2 つの素性 f_1、f_2 が独立の場合、素性の集合 F に f_1 を加えて確率モデルを推定した後に f_2 を加えた場合の対数尤度の増分 $\Delta L'(f_2)$ と、素性の集合 F に f_2 を加えた場合の対数尤度の増分 $\Delta L(f_2)$ はほとんど差がない。これを利用すると、いくつかの素性を同時に F に追加できる。また白井らは素性関数が確率モデルの推定にどれほど有効かを示す素性効用という関数を定義することで、付け加えるべき素性関数を選択する手法を提案している [16]。その他、Berger らも興味深い提案をしている [2]。

ただし、現実的には素性集合は経験的に設定し、素性選択アルゴリズムを用いない場合も多い。本来、意味のない素性関数については、その対応する素性パラメータが小さくなるので実害はない。計算可能な程度の規模の素性集合を用意するのが現実的であろう。

6 自然言語処理への応用

最大エントロピー法の自然言語処理への応用は多岐にわたる。なぜなら、最大エントロピー法は分類問題に対する手法であり、分類問題に変換できる様々な自然言語処理の問題は、当然、最大エントロピー法で解決できるからである。ここでは、特に成功したと思われる応用として、構文解析と固有表現抽出について紹介する。

英語の構文解析については Ratnaparkhi の最大エントロピー法に基づく学習モデルを利用した解析システム [9] が、精度、速度の両面で最も進んでいる手法と考えられている。そこでは構文解析が出力する木を、その木を構築するアクションの列 $\{a_1, a_2, \cdots, a_n\}$ によって表現する。あるアクションの列 $\{a_1, a_2, \cdots, a_k\}$ は構文解析の途中の段階を示し、この段階で次のアクション a_{k+1} を推定するのに最大エントロピー法を利用している。アクションは TAG、CHUNK、BUILD、CHECK の 4 つの手続きから生成される。そしてそれぞれに対する確率モデル $P_{TAG}(a|b)$、$P_{CHUNK}(a|b)$、$P_{BUILD}(a|b)$ および $P_{CHECK}(a|b)$ を最大エントロピー法により学習する。構文解析は 3 段階にわかれており、第 1 段階で TAG の処理を行い、第 2 段階で CHUNK の処理を行い、第 3 段階で BUILD と CHECK の処理を行う。

日本語の構文解析は係り受け解析である。係り受け解析においても最大エントロピー法を用いた内元のシステム [12] が現在最も精度がよいとされている。係り受け解析は、一般に、二文節間の係りやすさを数値化した係り受け行列を作成し、動的計画法などを用いて一文全体が最適な係り受け関係になるような係り先の組を求めることで行える。後半の問題は自然言語とは独立の探索問題であるため、実質的に問題となるのは、前半の部分、つまり二文節間の係りやすさをどのように求

めるかである。内元のシステムでは、二文節間の係りやすさを問題の二文節が係る確率と考えて、それを最大エントロピー法から求めている。二文節間の文脈情報 h に対して、その二文節が「係る(1)」と「係らない(0)」の2つのクラスを用意する。最大エントロピー法により $P(1|h)$ が求まる。

固有表現抽出も最大エントロピー法が成功した応用例である。固有表現抽出とは人名や会社名などの固有表現をテキストから抽出することであり、情報抽出の前処理として必要な処理である。固有表現抽出も分類問題に変換できる。例えば、人名を抽出するには、入力文の各単語に以下の5種類のクラスを割り当てればよい。

OP-CL : その単語自身が人名
OP-CN : 人名が複合語でその最初の単語
CN-CN : 人名が複合語でその中間の単語
CN-CL : 人名が複合語でその最後の単語
none : その単語は固有表現とは無関係

関根は8タイプの固有表現を扱っているので、合計32種類のクラスと none というクラスの計33種類のクラスを用意した。そして決定木を利用して、各単語に割り当てられるクラスの確率を求め、次に Viterbi アルゴリズムによって、尤もらしいクラスの列を生成した [11]。その決定木の部分に最大エントロピー法を用いたものが、Borthwick のシステム [4] である。このシステムは MENE と名付けられ MUC-7 で利用された [5]。また IREX において固有表現抽出コンテストが行われたが、そこでも、Borthwick[3] や内元 [14] が最大エントロピー法を用いて、優秀な成績を納めている。特に Borthwick のシステムは、IREX の NE 学習システムの中では最も良い成績を出している。

7 ツールの利用

機械学習手法自体は個々の問題とは独立であるため、ある機械学習手法を実装すれば、そのプログラムを再利用することが可能である。そのため機械学習手法を実装したプログラムを、ツールとしてパッケージ化したものが存在する。このようなツールは自然言語処理のような応用研究にとって有益である。

決定木に対しては C4.5 が有名である。Quinlan の書籍 [8] には C4.5 のアルゴリズムの説明、C言語によるコードが記されている。またそのコードを納めた FD も添付されており、自由に C4.5 が利用できる。このため、決定木を利用した応用研究の多くは、そこで提供されているパッケージを利用している。C4.5 は更に機能が付加されて C5.0 として販売されている。

http://www.rulequest.com

また決定木学習アルゴリズムでは CART も有名である。これは

http://www.salfordsystems.com/html/product.html

から購入できる。

Ristad は Maximum Entropy Modeling Toolkit という最大エントロピー法のパッケージを以下のアドレスで配布した。

http://www.mnemonic.com/software/memt

しかし現在すでに閉鎖されており、新たに上記のパッケージを入手することはできない。また、再配布も禁止されているので、すでに手にいれている人からもコピーして使うこともできない。残念である。別の最大エントロピー法のパッケージとしては、GIS アルゴリズムを Perl で記述したものも存在する。これは以下の URL を辿ることで入手できる。ただし、機能は低いと思われる。

http://www.sultry.arts.usyd.edu.au/links/statnlp.html

Ristad のツールのような最大エントロピー法のツールが、再びどこかで配布されることを期待したい。

8 おわりに

本稿では実装上の手引きになることを念頭に最大エントロピー法を解説した。最大エントロピー法を自然言語処理の分類問題に適用する場合、従来の決定木などの手法よりも精度がよくなる傾向がある。そのために最大エントロピー法は今後ますます利用されるであろう。本稿が最大エントロピー法を利用する際の参考になれば幸いである。

最後に最大エントロピー法の問題を２点述べておく。１点目は素性パラメータの推定に膨大な計算機資源が必要な点である。例えば前述した Ristad のツールは１ギガバイトのメモリを要求すると聞く。最大エントロピー法を手軽に試すためには、現状よりももう少し計算機環境が向上するか、アルゴリズムが改良される必要があろう。２点目は訓練データの問題である。これは帰納学習全般に言える問題であるが、訓練データを作成するコストが高いという問題である。自然言語処理の場合、事例は豊富にあるが、正解を付与するコストが大きい。近年、正解のない訓練データからの学習として、教師なし学習が注目されているが、精度に問題がある場合が多い。自然言語処理では訓練データ作成の問題も含めての問題解決が望ましい。

参考文献

[1] A. Berger, S. Della Pietra, and V. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[2] A. Berger and H. Printz. A Natural Criterion for Maximum Entropy / Minimum Divergence Feature Selection. In *3rd conference of Empirical Methods in Natural Language Processing (EMNLP-3)*, pp. 97–106, 1998.

[3] A. Borthwick. A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese. IREX ワークショップ予稿集, pp. 187–193, 1999.

[4] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting Diverse Knowledge Source via Maximum Entropy in Named Entity Recognition. In *6th Workshop on Very Large Corpora (WVLC-6)*, 1998.

[5] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. NYU: Description of the MENE named entity system as used in MUC-7. In *7th Message Understanding Conference*, 1998.

[6] J.N. Darroch and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470–1480, 1972.

[7] S Della Pietra, V. Della Pietra, and J. Lafferty. Inducing Features of Random Fields. *IEEE Transactions Pattern Analysis and Machine intelligence*, Vol. 19, No. 4, 1977.

[8] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publisher, 1993.

[9] A. Ratnaparkhi. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. In *2nd conference of Empirical Methods in Natural Language Processing (EMNLP-2)*, 1997.

[10] A. Ratnaparkhi. Natural Language Learning with the Maximum Entropy Framework. In *Tutorial of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99 Tutorial)*, 1999.

[11] S. Sekine, R. Grishman, and H. Shinnou. A Decision Tree Method for Finding and Classfying Names in Japanese Texts. In *6th Workshop on Very Large Corpora (WVLC-6)*, 1998.

[12] K. Uchimoto, S. Sekine, and H Isahara. Japanese Dependency Structure Analysis Based on Maximum Entorpy Models. In *9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pp. 196–203, 1999.

[13] D. Yarowsky. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *32th Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 88–95, 1994.

[14] 内元清貴, 村田真樹, 小作浩美, 馬青. ME モデルと書き換え規則に基づく固有表現抽出 - IREX-NE 本試験における評価-. IREX ワークショップ予稿集, pp. 133–140, 1999.

[15] 白井清昭. 統計情報を利用した統合的自然言語解析. 博士論文. 東京工業大学, 1998.

[16] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 最大エントロピー法による確率モデルの推定に有効な素性の選択について. 言語処理学会第 4 回年次大会, pp. 356–359, 1998.

[17] 北研二. 確率的言語モデル. 東京大学出版会, 1999.