最大エントロピー法の解析的解法

掛谷英紀

筑波大学システム情報工学研究科

概要　最大エントロピー法による機械学習は、スパースなデータに対応できることから、自然言語処理にしばしば使われる手法である。最大エントロピー法におけるパラメータ推定では、繰り返し演算による数値的解法が知られているが、学習対象のデータサイズが大きくなると、学習に要する時間が増大するという問題がある。そこで、本研究では、自然な制約条件を新たに導入することで、最大エントロピー分布を解析的に求める方法を提案する。さらに、提案した方法を用いて実際に学習を行い、文書分類の問題における従来手法とのパフォーマンスの違いを比較する。

1　背景

文書分類において、最大エントロピー法 [1, 2] はサポートベクターマシンなどと並んで、頻繁に使われる機械学習法の一つである。最大エントロピー法の利点は、それぞれの素性の判定における寄与率を数値化できる点にある。その一方で、学習（パラメータ推定）には繰り返し演算による数値的解法を用いるため、学習対象のデータサイズが大きくなると、学習に要する時間が増大するという問題もある。

筆者らは、これまで、最大エントロピー法を用いた機械学習を使って、大手新聞社の社説や国会議事録における各政党の発言を題材にして最大エントロピー法に基づく機械学習を行い、そのイデオロギー的特徴を分析してきた [3, 4, 5]。しかしながら、過去１０年の国会議事録全てについて、単語や熟語だけでなく単語や熟語の共起なども素性として取り扱う場合、学習に多大な時間を要するという問題があった。

そこで、本研究では、文書分類問題において、自然な制約条件を新たに導入することで、最大エントロピー分布を解析的に求める方法を提案する。さらに、提案した方法を用いて実際に学習を行い、文書分類の具体的問題について従来手法とのパフォーマンスの違いを比較する。

2　最大エントロピー法

文書分類においては、それぞれの文書に含まれる素性のリストから、どのカテゴリに属するかを判定する形をとるのが一般的である。その一つの方法として、各素性の組み合わせが出現する頻度をもとに確率分布を作り、その分布からベイズ推定を行うことが考えられる。しかし、文中に含まれる素性候補は一般に極めて多く、数万から数十万のオーダーになることも少なくない。素性数が n であれば、素性の出現組み合わせは 2^n 通りあり、これは学習データ数 m を遙かに上回るのが普通である。よって、学習データの出現頻度を忠実に再現する確率分布では、確率のほとんどが 0 または $1/m$ になり、実用的に意味をなさない。このように、非常にスパースなパターンについて有意な確率分布を与える方法の一つが最大エントロピー法である

今、カテゴリ数を p とし、それぞれのカテゴリを $a_1, a_2, \cdots, a_p$ と表わすことにする。一方、素性に $b_1, b_2, \cdots, b_n$ と番号をふることにする。ここで、ある文書に k 番目の素性が含まれている場合は $b_k = 1$、含まれていない場合は $b_k = 0$ と表わすことにすると、文書はベクトル $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$ で表わすことができる。最大エントロピー法では、素性パターン $\boldsymbol{b}$ をもつカテゴリ a_i の文書の出現確率 $P(a_i, \boldsymbol{b})$ を、次の (A1)、(B) の仮定に基づき推定する。

(A1) 学習データにおいて $b_k = 1$ を満たすカテゴリ a_i の文書数を f_{ik}、それを全文書数で割った量を $q_{ik} = f_{ik}/m$ と定義したとき、

$$q_{ik} = \sum_{\{\boldsymbol{b}|b_k=1\}} P(a_i, \boldsymbol{b}) = \sum_{\boldsymbol{b}} b_k P(a_i, \boldsymbol{b}) \qquad (1)$$

が成り立つと仮定する。これは、導出する確率分布と学習データの間で、カテゴリ a_i の文書中に各素性が出現する期待値が等しいことを意味する。

(B) 確率分布 $P(a_i, \boldsymbol{b})$ のエントロピー

$$-\sum_i \sum_{\boldsymbol{b}} P(a_i, \boldsymbol{b}) \log P(a_i, \boldsymbol{b}) \qquad (2)$$

が、(A1) を満たすという条件下で最大値をとると仮定す

る。これは、(A1) の条件が満たされる範囲で、確率分布は最大限なだらかになっていることを意味する。

上記の条件を満たす確率分布は、上述の通り数値的に求める方法が知られている。

エントロピーを最大にする分布が求まれば、素性パターン $\boldsymbol{b}$ を持つテストデータが与えられたとき、その文書がカテゴリ a_i に属する確率は

$$P(a_i|\boldsymbol{b}) = P(a_i, \boldsymbol{b}) / \sum_j P(a_j, \boldsymbol{b}) \tag{3}$$

と計算される。明確な判別が必要な場合は、$P(a_i|\boldsymbol{b})$ の値が最も大きくなるカテゴリ a_i に判別すればよい。

3 最大エントロピー法の解析的解法

条件が上記の (A1) の仮定だけの場合、エントロピーを条件付きで最大にする確率分布を解析的に求めることはできない。そこで、条件 (A1) に加え、新たに次の仮定 (A2) を条件として加えることを考える。

(A2) 学習データにおけるカテゴリ a_i の文書数を f_i とし、それを全文書数で割った量を $q_i = f_i/m$ と定義したとき、

$$q_i = P(a_i) = \sum_{\boldsymbol{b}} P(a_i, \boldsymbol{b}) \tag{4}$$

が成り立つと仮定する。これは、導出する確率分布と学習データの間で、各カテゴリの文書が出現する期待値が等しいことを意味する。

以下、表記を簡単にするため、次のような変数を導入する。まず、ありうる 2^n 個の素性パターン $\boldsymbol{b}$ に通し番号をふり、j 番目の素性パターン $\boldsymbol{b}^{(j)}$ と表す。カテゴリ a_i の素性パターン $\boldsymbol{b}^{(j)}$ が出現する確率 $P(a_i, \boldsymbol{b}^{(j)})$ を X_{ij} と表記すると、この確率分布のエントロピー H は

$$H = -\sum_{i=1}^{p}\sum_{j=1}^{2^n} X_{ij} \log X_{ij} \tag{5}$$

で与えられる。一方、制約条件 (A1) は

$$\sum_{j=1}^{2^n} b_k^{(j)} X_{ij} = q_{ik} \tag{6}$$

と表わすことができ ($i = 1, 2, \cdots, p, k = 1, 2, \cdots, m$)、制約条件 (A2) は

$$\sum_{j=1}^{2^n} X_{ij} = q_i \tag{7}$$

と表わされる ($i = 1, 2, \cdots, p$)。ここで、$b_k^{(j)}$ は j 番目の文書 $\boldsymbol{b}^{(j)} = (b_1^{(j)}, b_2^{(j)}, \cdots, b_n^{(j)})$ において k 番目の素性が含まれているか否かを表す（含まれていれば 1、含まれていなければ 0）。また、q_{ik} は学習データにおける i 番目のカテゴリで k 番目の素性を含むパターンの出現確率、q_i は学習データにおける i 番目のカテゴリの出現確率である。

この制約条件を満たしつつ、最大のエントロピーを与える X_{ij} は、ラグランジュの未定乗数 λ_{ik}, λ_i を用いると、

$$\begin{aligned} -\frac{\partial H}{\partial X_{ij}} \quad &+ \quad \sum_k \lambda_{ik} \frac{\partial}{\partial X_{ij}} (\sum_{j=1}^{2^n} b_k^{(j)} X_{ij} - q_{ik}) \\ &+ \quad \lambda_i \frac{\partial}{\partial X_{ij}} (\sum_{j=1}^{2^n} X_{ij} - q_i) = 0 \end{aligned} \tag{8}$$

を満たすことになる ($i = 1, 2, \cdots, p, j = 1, 2, \cdots, 2^n$)。この偏微分を計算すると

$$-(1 + \log X_{ij}) + \sum_k \lambda_{ik} b_k^{(j)} + \lambda_i = 0 \tag{9}$$

となり、

$$X_{ij} = \exp[\sum_k \lambda_{ik} b_k^{(j)} + \lambda_i - 1] \tag{10}$$

が得られる。ここで、λ_{ik}, λ_i を求めることができれば、$X_{ij} = P(a_i, \boldsymbol{b}^{(j)})$ が求まり、所望の確率分布が得られる。そこで、新たに変数 $Y_{ik} = \exp[\lambda_{ik}]$ と $C_i = \exp[\lambda_i - 1]$ を導入すると、上の式は

$$X_{ij} = C_i \prod_{b_k^{(j)}=1} Y_{ik} \tag{11}$$

と表すことができる。これを用いると、制約条件 (A1) に対応する式 (6) は

$$C_i \sum_{j=1}^{2^n} b_k^{(j)} \prod_{b_k^{(j)}=1} Y_{ik} = q_{ik} \tag{12}$$

と表わすことができる ($k = 1, 2, \cdots, n$)。この式を因数分解すると、各素性 b_t について

$$C_i Y_{it} \prod_{k \neq t} (1 + Y_{ik}) = q_{it} \tag{13}$$

の関係式が得られる。また、制約条件 (A2) に対応する式 (7) に式 (11) を代入し、因数分解すると

$$C_i \prod_k (1 + Y_{ik}) = q_i \tag{14}$$

も成り立つことが分かる。ここで、式 (13) と式 (14) を比較すると

$$Y_{it}/(1 + Y_{it}) = q_{it}/q_i \tag{15}$$

が成り立つことから

$$Y_{it} = q_{it}/(q_i - q_{it}) \tag{16}$$

が得られる。よって、所望の確率分布は

$$P(a_i, \boldsymbol{b}^{(j)}) = X_{ij} = C_i \prod_{b_k^{(j)}=1} q_{ik}/(q_i - q_{ik}) \tag{17}$$

で与えられることになる。ここで、C_i は $\sum_j X_{ij} = q_i$ とする規格化定数に相当することから

$$C_i = \frac{q_i}{\sum_{j=1}^{2^n} \prod_{b_k^{(j)}=1} q_{ik}/(q_i - q_{ik})} \tag{18}$$

で与えられる。

なお、ラグランジュの未定乗数法を直接解いた上記のパラメータは、この確率モデルの最尤推定量になっている。この確率モデルで、学習データと同じ出現頻度 q_{ik} が生じる確率 $P(q_{ik})$ は、

$$P(q_{ik}) = \prod_i \left(C_i^{q_i} \prod_k Y_{ik}^{q_{ik}}\right) \tag{19}$$

で与えられる。この対数をとると、

$$\log P(q_{ik}) = \sum_i q_i \log C_i + \sum_i \sum_k q_{ik} \log Y_{ik} \tag{20}$$

となり、これが対数尤度となる。この対数尤度を、確率の総和が 1 であるという条件

$$\sum_i C_i \prod_k (1 + Y_{ik}) = 1 \tag{21}$$

のもとで最大化するパラメータが最尤推定量になる。再びラグランジュの未定乗数法を用いると、

$$\begin{aligned} L(C_i, Y_{ik}) &= \sum_i q_i \log C_i + \sum_i \sum_k q_{ik} \log Y_{ik} \\ &+ \lambda \left(\sum_i C_i \prod_k (1 + Y_{ik}) - 1\right) \end{aligned} \tag{22}$$

を各パラメータで偏微分した式が 0 になることが極値をとる条件となる。よって、

$$\frac{\partial L}{\partial C_i} = \frac{q_i}{C_i} + \lambda \prod_k (1 + Y_{ik}) = 0 \tag{23}$$

$$\frac{\partial L}{\partial Y_{it}} = \frac{q_i k}{Y_{it}} + \lambda C_i \prod_{k \neq t} (1 + Y_{ik}) = 0 \tag{24}$$

が満たされる C_i, Y_{ik} が最尤推定量となる。この連立方程式を解くと、その解はラグランジュ未定乗数法を直接解いて求めた式 (17),(18) と等しくなる。

判別においては、$\boldsymbol{b}^{(j)}$ がカテゴリー a_s に属する確率は、

$$P(a_s|\boldsymbol{b}^{(j)}) = \frac{C_s \prod_{b_k^{(j)}=1} q_{sk}/(q_s - q_{sk})}{\sum_{i=1}^{p} C_i \prod_{b_k^{(j)}=1} q_{ik}/(q_i - q_{ik})} \tag{25}$$

で与えられることになる。

判別結果を確率分布としてではなく、一意に出力したい場合は、$P(a_s|\boldsymbol{b}^{(j)})$ の値が最も大きくなるカテゴリ a_s であると判定すればよい。

4 実験

前節で導いた解析的手法の実用性を検証するため、実際の文書分類問題において、本手法が従来手法と同等の判定精度を達成できるかどうかを試す実験を行った。問題としては、橋本らによる体験談捏造判定問題を使用する [6]。

最近、商品広告において、嘘の体験談をライターに書かせているケースがしばしば見られる。アガリクスの癌に対する効能の体験談を大量に捏造・出版し、その後薬事法違反で摘発された史輝出版はその代表的事例である。こうした偽の体験談に基づく広告は、不正が発覚していないものの中にも多く含まれている可能性がある。

橋本らは、同種の商品の体験集を広告サイトごとに比較し、ある広告サイトの体験集が他のサイトの体験集と機械学習で区別できる場合、同一ライターによって書かれたために、そのライターの癖が手がかりとなったもので、捏造の可能性が高いと推定する方法を提案している。もちろん、多数のライターを使っての捏造については、この方法では判定できないが、多人数がコミットした不正はその発覚確率を高める恐れがあることから、この方法による捏造判定は一定の妥当性があると考えられる。

橋本らは、バイアグラと精力剤の広告を具体例として、２つの同一のサイトから集めた体験集Ａと体験集Ｂ、および複数のサイトから集めた体験集Ｃ（いずれも約１００件）について最大エントロピー法を用いた機械学習を試み、体験集Ａは体験集Ｂ、Ｃと高い精度で判別可能である一方、体験集Ｂと体験集Ｃは十分に判別できないことから、体験集Ａは同一ライターによるフィクションの可能性が高く、体験集Ｂは実際に複数の体験談を収集したものである可能性が高いと結論づけている。

今回は、橋本らが従来の最大エントロピー法の学習プログラム maxent[7] を用いて行った上記の実験結果を、前節の解析的手法で再現できるかどうかを試みる実験を行った。

前節の方法を具体的問題に適用する場合、２点細かな修正が必要である。まず、第一点は $q_i = q_{ik}$ の場合、すなわちあるカテゴリで全ての文書で出てくる素性がある場合、式が発散する問題である。これについては、全ての文書で

出現する素性は q_i 回ではなく $q_i - \delta$ 回出現しているとして処理すれば対応可能である $(0 < \delta < 1)$。次に、第二点として、あるカテゴリでどの文書にも出てこない素性がある場合、その素性を含む文書の出現確率が 0 になってしまう問題がある。これは実用上強すぎる制約なので、どの文書にも出現しない素性は 0 回ではなく ϵ 回出現しているとして処理することにする $(0 < \epsilon < 1)$。

橋本らの実験結果と、同じデータを本手法で判別させた結果を比較したものを表 1 に示す。この表が示すとおり、$\delta = \epsilon = 0.5$ の場合も、$\delta = \epsilon = 0.1$ の場合も、maxent とほぼ同様の正解率の傾向を示していることが確認された。

なお、学習に要する時間は、解析的にパラメータが導出できるようになったため、繰り返し学習が必要な maxent に比べると飛躍的な短縮が見られている。

表 1: 各学習方式での正解率の比較

	maxent	提案方式 $(\delta, \epsilon : 0.5)$	提案方式 $(\delta, \epsilon : 0.1)$
AとBの判別	90.5 %	89.5 %	90.5 %
AとCの判別	91.7 %	88.7 %	89.2 %
BとCの判別	60.5 %	67.6 %	66.7 %

5 まとめ

本研究では、文書分類問題において、自然な制約条件を新たに導入することで、最大エントロピー分布を解析的に求める方法を提案した。提案した方法を用いて、同一ライターによる体験談捏造検出問題について、従来手法と比較を行い、ほぼ同様の結果を与えることが確認されると同時に、パラメータ推定に要する時間を大幅に短縮できることを確認した。

今回用いた例題は、学習データのサイズが比較的小さな問題である。今後は、計算量低減の恩恵が大きい、より大きなサイズの学習データを対象とする文書分類問題について、提案手法の有効性を検討する予定である。また、δ と ϵ の設定方法に関する理論的、実験的考察も今後の検討課題である。

参考文献

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. (1996) A Maximum Entropy Approach to Natural Language Modeling. Computational Linguistics, 22(1).

[2] Adam Berger. (1997) The Improved Iterative Scaling Algorithm: A Gentle Introduction, Technical Report, CMU.

[3] 木村弦, 金丸敏幸, 村田真樹, 掛谷英紀 (2007) 新聞の社説を教師信号とする文章の右翼度・左翼度判定, 言語処理学会第 13 回年次大会講演論文集.

[4] 畑中允宏, 金丸敏幸, 村田真樹, 掛谷英紀 (2008) 新聞の社説を教師信号とする文章の右翼度・左翼度判定第二報, 言語処理学会第 14 回年次大会講演論文集.

[5] 畑中允宏, 村田真樹, 掛谷英紀 (2009) 新聞社説・国会議事録に基づく言論のイデオロギー別分類, 言語処理学会第 15 回年次大会講演論文集.

[6] 橋本悠, 掛谷英紀 (2009) 自然言語処理による広告及び経営トップのメッセージ分析, 第 5 回メディア情報検証学術研究会講演論文集.

[7] The openNLP MAXENT package. http://maxent.sourceforge.net/.