モジュール学習 — Finite Mixture と EM —

赤穂 昭太郎 (PY)
電子技術総合研究所

Modular Learning — Finite Mixture and EM —

Shotaro Akaho <akaho@etl.go.jp>
Electrotechnical Laboratory

Abstract — In this paper, we review some theoretical results for modular learning. Modularity is effective especially to solve complicated problems, and mixture modelling is a simple implementation of modularity. In oder to train the mixture models, the EM algorithm is often applied to optimize model parameters because of its simplicity and robustness.

1. はじめに

一見複雑に見える問題も, 問題の空間を適切に分割することによって易しくなることがある. ニューラルネットの場合も, 一つの大きなネットワークで対象の全体を学習するのではなく, 全体を小さなモジュールに分割し, 各モジュールが分担して学習を行なって, それらを全体として統合することにより学習が容易になったり汎化能力が向上することが期待される (モジュール学習).

従来そのようなアプローチは分割統治法として知られている. 関数近似におけるスプラインなども局所的に単純な関数の組合せで全体としては複雑な関数を実現するという意味で, その最も単純な場合であるということができる.

単純な分割統治では分割の仕方はあらかじめ決まっているのが普通である. しかし高次元の複雑な問題を扱う場合に, どのように問題を分割してよいかがわからない場合も多い (セグメンテーションの問題). ここでは, 分割の仕方と各モジュールの能力の両方を学習するような枠組について考察し, その理論的・実用的課題について述べる.

2. Finite Mixture

ここではモジュール学習の基本となる確率モデルを導入する. ニューラルネットにおいてもそうだが, 確率モデルとして扱うことによって学習の解析をすることが容易になる.

学習には大きく分けて, 教師なし学習と教師あり学習とがある. 教師なし学習はデータ x の確率分布 $p(x;\theta)$ の推定であり, 教師あり学習は入力データ x から出力 y への条件つき確率分布 $p(y \mid x;\theta)$ の推定であるとみなすことができる.

Keywords — EM algorithm, Finite Mixture Model, Modular Learning

さて, 教師なし学習を行なう K 個のモジュール $p_k(x;\theta_k), k = 1,\ldots,K$ があったとする. 確率 ξ_k で k 番目のモジュールのデータが得られるとすると全体の確率モデルは,

$$p(x;\theta) = \sum_{k=1}^{K} \xi_k \, p_k(x;\theta_k), \tag{1}$$

となる (ただし一般に k は与えられない). 同様に, 教師ありの場合も, モジュール $p_k(y \mid x;\theta_k)$ のデータが確率 $\xi_k(x;\psi_k)$ で得られるとしたとき (x に依存してよい), 全体の確率モデルは

$$p(y \mid x;\psi,\theta) = \sum_{k=1}^{K} \xi_k(x;\psi_k) \, p_k(y \mid x;\theta_k), \tag{2}$$

となる. (1) や (2) のことを (有限) 混合分布モデル (Finite Mixture) という. 以下にそれぞれの典型的な例を挙げる.

正規混合分布 各モジュールが正規分布

$$p_k(x;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}, \tag{3}$$

であるような教師なしのモデルである. 容易にわかるように, $\xi_k \neq 0$ であって $\sigma_k^2 = 0$ となると, このモデルの尤度は無限大になる. これが最尤法におけるグローバルな最適解であり, ただ一つのサンプルだけに一つのモジュールがフィットしてしまう場合に起こり得るが, 明らかにこれは求める解ではない. 従ってこのような無意味な解に収束し難いような安定なアルゴリズムが必要となる.

エキスパート混合モデル Jordan ら [6] は, 教師あり学習 (2) の例として,

$$\xi_k(x;v_k) = \frac{\exp(v_k x)}{\sum_{j=1}^{K} \exp(v_j x)}, \tag{4}$$

$$p_k(y \mid x;w_k) = \frac{1}{1+\exp(w_k x)}, \tag{5}$$

なるモデルを提案した (y=0,1). さらに, これを階層化することにより, (計算法は簡単なまま) より複雑なモデルを作ることができる (Hierarchical Mixtures of Experts). Amari[4] は階層化ではないより簡単な方法で, 同様の複雑化を行なう方法を提案している.

3. EM アルゴリズムと幾何学的解釈

モジュール学習を行なうためのアルゴリズムはいろいろなものが考えられる. 基本的には非線形の最適化問題であるから, Newton 法や最急勾配法などを適用することもできる. しかしながら, 混合モデルに対しては EM アルゴリズムを自然な形で適用できるため, 本稿では EM アルゴリズムに限って説明を行なう.

EM アルゴリズム [5] は, 最尤推定を求めるための反復法の一つである. 定義は本節の最後で述べることにし, まずは幾何学的観点からその仕組みの概略を説明する [4, 2] (図 1).

確率分布はそのパラメータを局所座標系とする多様体とみなすことができる. ある空間 S を考えると, 与えられたデータ x は S の中の 1 点 $r(x)$ として表すことができ, 推定する確率モデルは S の部分多様体 M として表すことができる. 最尤推定は, データ $r(x)$ からモデル M への 射影であるととらえることができる. もし M が平らな空間ならば射影は易しいが, 曲がりくねっていると難しくなる (図 1 左).

EM アルゴリズムでは, x は背後にある完全なデータ z の不完全なデータであるとみなすことによって, 推定の問題を z の空間での推定に帰着させることを考える. もし, z の空間で考えることによってモデル M が平らになるとすれば問題は易しくなるであろう.

混合モデルに関して言えば, どのモジュールのデータであるかという k を観測できない変数であると考えることによって, $z = (k, x)$ の確率分布を考えることに相当する. もし, k がわかっていればモジュールの学習は格段に易しくなる.

ただし, 不完全なデータ x は z の分布の空間ではもはや 1 点ではなく, 不完全な分の自由度を持った多様体になる. 従って, 問題は多様体と多様体の間の距離が最短となるような点を見つけることになる (図 1 右). Amari[4] は, データの多様体とモデルの多様体の間で射影を繰り返すことによりその局所最適解を見つける方法として em アルゴリズムというものを提案した. 実はこの em アルゴリズムと EM がほとんどの場合に一致することが示されている. ただし, (特にサンプル数が少ない場合には) 異なる場合が存在し, どちらがより自然なものかが議論されている.

上の説明では実際の計算がどうなっているかがわからないので, EM アルゴリズムの定義を述べる. EM アルゴリズムの各ステップは z の対数尤度を最大化するが, 観測できない部分については, 現在のパラメータ θ^t と与えられたデータ x に対して条件つき期待値をとる.

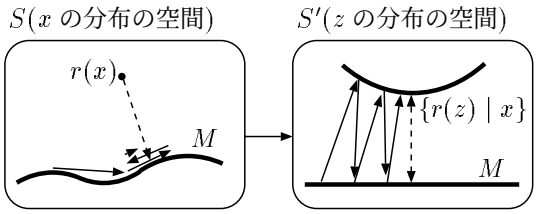

図 1. EM アルゴリズム

式で書けば,

$$\theta^{t+1} = \operatorname*{argmax}_{\theta} \mathrm{E}_z[\log p(z;\theta) \mid x, \theta^t], \qquad (6)$$

ということになる. 期待値をとる部分を E(xpectation) step, 最大化の部分を M(aximization) step と呼ぶが, モデルによっては One step で最大化までできてしまって分離できないこともある.

4. EM アルゴリズムの収束性

EM アルゴリズムの最も重要な性質は, 各繰り返しステップにおいて尤度が単調に増加することである (これによって一般には按点に収束する). Newton 法などでは必ずしもこの性質は満たされない. 従って, 比較的よい初期値から始めれば, 安定に局所解を求めることができ, 尤度無限大という無意味な解が存在する正規混合分布などに適している.

収束の速度に関しても, 特に繰り返しの初期はかなり速いことが知られている. Newton 法の近似であることを示す結果もある.

ただし, 局所最適解 θ^* に近付いて来ると一次の収束しかしないことが示されている. 収束の様子を一次近似すると,

$$\theta^{t+1} - \theta^* \simeq [I_{\mathrm{com}}^{-1} I_{\mathrm{mis}}](\theta^t - \theta^*), \qquad (7)$$

という結果が得られる. ここで, I_{mis} は, 観測できない部分の Fisher 情報量の観測データに対する条件つき期待値であり, I_{com} は完全データの Fisher 情報量の条件つき期待値である. ユークリッドノルムで距離を考えると, 収束の速さは右辺の [ ] 内の行列の最大固有値の大きさで測ることができる (大きいほど収束は遅い). この式は, 欠損した情報量が多いほど収束が遅くなるという直観的な結果を表す式になっている. モジュール学習においても, 正規混合分布やエキスパート混合モデルについての具体的な計算がなされている.

まとめると, EM アルゴリズムは粗い安定した解を求めるのに十分速い収束性を示すが, さらに精度を上げるためには Newton 法と組み合わせたりいろいろ知られている数列の加速法を用いたりする必要がある (収束が安定しているのでこのような組み合わせは比較的相性がよい).

5. インプリメンテーションの容易さ

EM アルゴリズムが用いられるもう一つの理由はインプリメンテーションの容易さである. Newton 法のよ

うに微分した関数の逆行列を計算する必要がない. ただし, (6) 式は一般には最大化の部分が残っているので, Newton 法なり最急勾配法なりを用いて解かねばならないが, z の分布が指数型分布族

$$p(z;\theta) = \exp\{\theta F(z) - \psi(\theta) + C(z)\} \qquad (8)$$

の場合には

$$\eta^{t+1} = \mathrm{E}_z[F(z) \mid x;\theta^t] \qquad (9)$$

という条件付き期待値の計算だけで E step + M step が終わってしまう. ここで η というのは θ に双対なパラメータで, $\eta = \mathrm{E}_z[F(z);\theta]$ または $\eta = \partial\psi(\theta)/\partial\theta$ で定義される. η^{t+1} から θ^{t+1} へは単純な座標変換で移すことができる.

6. モジュール学習の EM

モジュール学習において, z の分布が指数分布族に属するのは教師なしでは正規混合分布などモジュールが指数型分布属の場合であり, 教師ありではエキスパート混合モデルの線形の場合などの単純なモデルに限られる. 指数型分布族には属さないが, 陽に解ける例として正規混合分布で作った分布をテンプレートとして, スケールとシフトのパラメータをもつモジュールをもつ教師なしの混合モデル [3] がある. 非線形な教師あり学習の場合には一般に (複数サンプルの空間では) 指数分布族には属さないが, エキスパート混合モデルは一般化線形モデルと呼ばれる形を採用しているので, かなり安定した反復推定法が確立している.

このように指数分布族に属すものがあったりなかったりするが, どちらの場合においても, モジュール学習において EM を用いるもう一つのメリットは処理の独立性あるいは局所性である. 各モジュールの学習においては他のモジュールの詳細な構造は必要がない. 具体的には各モジュールは, グローバルな情報として, 各サンプルに対する $p(x;\theta)$ という量だけを必要とするが, これは local な情報の単純な和であり, 簡単に計算できる. もちろん学習はモジュールの複雑さに応じて複雑になる可能性はあるが, 他のモジュールにその複雑さの影響を及ぼすことはない. このことは分割統治を効率的に行なえることを示すと同時に, モジュール学習においていろいろなタイプのモジュールを組み合わせて使うことの容易さをも意味している.

7. 計算上の問題点

EM アルゴリズムはモジュール学習以外にも, Bayes モデルにおけるハイパーパラメータの推定に用いることができるなど幅広い推定に適用可能なアルゴリズムである. そのため, あてはめるモデルによっては計算が大変になることがある.

まず, E step は条件付きの期待値を計算するため, 数多くの項の和を計算しなければならなかったり, 場合によっては数値積分を必要とする場合がある. この部分の計算量を抑えるために近似的に用いられる方法は MCMC(Markov Chain Monte Carlo) 法や平均場近似といった手法である. これらの近似手法によってどれだけ EM の性能が落ちるのか, あるいは逆に性能を保ったままどこまで近似してよいのかといった問題が提起される.

次に M step においては一般に非線形最適化を行なわなければならない (もとの問題よりは簡単であることは期待されるが). ただし, M step では最適解が求まらなくても, 最適化すべき関数の値を現在値より増加させさえすれば, 尤度の単調性は保証されるので, Newton 法などの勾配法を組み合わせて用いることが考えられる (一般化 EM, GEM). この Newton 法は必ずしも収束させるまでやる必要はない. また, 場合によっては, 変数をいくつかの組に分け, 他の変数は固定したままある変数の組を最適化していくというやり方がうまくいく場合もある (条件つき最大化による EM, ECM).

8. 局所解と初期化

EM アルゴリズムは局所最適解を求める手法であり, いくつかの望ましくない局所解に収束することがある. ここではモジュール学習の場合に典型的に起きる局所解について述べる.

まず, 一つ目のタイプは全てのモジュールが問題の空間全体に対してフィットしてしまうような場合である. つまり, 全く役割分担を行なわないため, モジュール学習のメリットが全然出ない場合である. 特に同一種類のモジュールを用いた場合にはすべてが同じモジュールに収束するような事態も起こり得る.

二つ目のタイプは学習途中でモジュールの現在の分担場所と, 望ましい分担場所とが全くオーバーラップしていない場合である. 定性的には, EM アルゴリズムでは望ましい分担場所に引っ張られるように収束していくので, オーバーラップがないと全く引っ張られないという現象が起きる.

三つ目のタイプは二つ目のタイプと関係しているが, 異なるモジュールを使った場合に, モジュール A で分担すべき部分がモジュール B で分担され, 逆にモジュール B で分担すべき部分がモジュール A で分担されてしまうという局所解があり, この場合, 正しい方向に収束させるのがかなり難しい.

これらの局所解を避けるためには, 通常のニューラルネットの学習の場合と同様に適切な初期解を与えてやることが必要である. もちろん完璧な初期解というのは存在しないが, 粗い近似解を与えることは重要である. このためのアプローチとしては構成的に初期ネットワークを構築する方法 [7] や, なんらかの heuristics や問題の構造に応じた事前知識を用いることによる方法などが検討されている.

9. モデル選択

パラメータの初期化も重要であるが, もっと根本的にモジュールの構造や個数を決める問題がある. EM アルゴリズムは与えられた構造のネットワークのパラ

メータを学習するだけなので, 自己組織的にモジュールの構造や個数を変化させるためにはメタなレベルでの操作が必要となる.

モジュール学習の場合も AIC などを使って, ニューラルネットの中間層を増やしたり減らしたりするのと同様のモデル選択を行なうことが考えられる. ただし, 大きい粒度のモジュールを扱うのはなかなか困難がある. 以下では, これとは多少違うやり方でモデル選択を行なう方法を, そこで見られる興味深い現象とともに紹介する [1].

ここで考えるモデルは単純な正規混合分布であり, 簡単のため ξ_k は定数 $1/K$ とするが, そのかわりモジュール数 K は非常に大きくとる (サンプル数と同程度のオーダーで). また, 正規分布の分散 σ_k^2 はすべて同じ固定の値 σ^2 とする. このモデルはボルツマンマシンの連続化と深い関係がある. σ^2 を決めるとモデルが一つ決まり, σ^2 を変えることがモデル選択に対応する.

σ^2 をいろいろな値に変えて最尤推定の結果を追跡してみると, σ^2 がある値 (サンプルの分散) 以上の間はすべてのモジュールはある 1 点 (サンプルの中心) に縮退する. その値を境に, モジュールが分岐現象を起こし, また σ^2 がある値以下になるとさらに分かれていくという具合いに, 階層的なクラスタリングと同様の構造を示す.

分岐点付近の最尤解を解析すると, (対称な分布については) 以下の 2 つの性質があることがわかった.

1. 分岐の仕方はサンプル分布の尖度によって分類でき, 正規分布と同じ尖度のときに限り連続的に分岐し, それ以外の場合には 2 つの方向に分かれ, その分かれ方は分散の軸に対して垂直である (図 2).

2. 2 つの方向に分かれる場合に, 汎化能力を示す基準となる有効パラメータ数 (NIC) を計算すると, 分岐点の前では NIC が単調増加するのに対し, 分岐点の後ろでは NIC が減少することが示せる. しかもほとんどすべての場合に傾き $-\infty$ で減少する (図 3).

特に, 第 2 の点は興味深い性質である. 分岐することによって見掛けのパラメータ数は増加するはずなのに実際は減少する. これに対する直観的な説明はまだないが, もしこういった性質がモジュール学習全般に普遍的であるとすると, モジュールを増やすことによって逆に汎化能力が上がったりすることもあり得るし, 逆にそういった非単調性がモデル選択を困難にするという危険性もはらんでいる.

10. まとめ

モジュール学習を混合分布の推定問題ととらえ, そのための推定アルゴリズムとして EM アルゴリズムに限って理論的な結果や問題点について概要を述べた. ニューラルネットの学習と同様に, 初期値や局所解の問題が存在するし, ネットワーク構造の決定やモジュールの設計など実用的に使うにはまだ難しい点がある. また

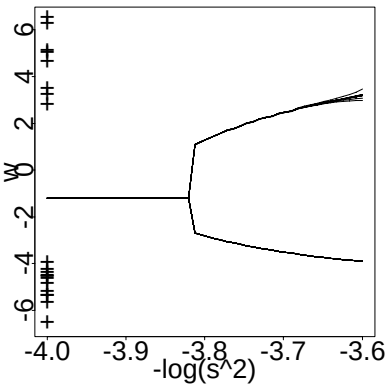

図 2. 分岐図

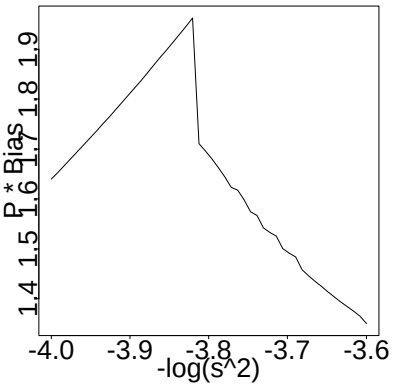

図 3. 尤度のバイアスの変化

収束性や汎化能力などに関しても, 今後更に理論的解析を行なう必要があると思われる.

References

[1] 赤穂: EM アルゴリズムの幾何学. 情報処理, Vol. 37, No. 1, pp. 43–51, 1996.

[2] 赤穂: 連続値ボルツマンマシンの分岐現象について. 信学技法, NC–95–92, 1996.

[3] Akaho, S.: The EM algorithm for multiple object recognition. In *Proc. of ICNN'95*, pp. 2426–2431, Perth, 1995.

[4] Amari, S.: Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, Vol. 8, No. 9, pp. 1379–1408, 1995.

[5] Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, Vol. 39, pp. 1–38, 1977.

[6] Jordan, M. I. and Jacobs, R. A.: Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, Vol. 6, pp. 181–214, 1994.

[7] 斉藤, 中野: HME の構成的学習アルゴリズム. 信学技法, NC–95–114, pp. 99–105.