隠れマルコフモデルにおける集合型パラメータ推定に関する研究

東京大学大学院情報理工学系研究科数理情報学専攻
黒澤 雅人 (学籍番号 48096215)
指導教員 中川 裕志 教授

1 はじめに

機械学習とは与えられたデータから, 未知のデータの性質や振る舞いを予測するアプローチである [1]. 例として音声認識, 自然言語処理等のタスクが挙げられる. これらのタスクにおいて, 教師あり学習の場合では, 機械的に得た生データに人手で正解を加えた教師データによって学習を行うが, 正解データを人手で作る必要があるため, 教師データの作成に対する人的コストが大きい. それに対し, 教師なし学習の場合は生データのみを与えて学習を行うため, 少ない人的コストで良い学習器を作れる可能性がある. しかし教師なし学習は, 過学習や複数の局所解の存在といった問題が多く起こる. この問題の対策として 3 つのアプローチが考えられる.

第 1 のアプローチは与えられたデータに対する新しいモデルの提案である. データに対してより柔軟なモデルを提案することで, 学習において起こりうる特異性を取り除き, 過学習を防ぐ役割を果たしている. 第 2 のアプローチは与えられたモデルとデータに対する新しい学習アルゴリズムの考案である. 元々存在する解の中でより良い解の探索を目指し, 局所解を防ぐ役割を果たしている. これらとは別に第 3 のアプローチとして, 与えられたモデルに対するより良い予測手法の考案がある. 例としてはアンサンブル学習や正則化である. この第 3 のアプローチは第 1, 第 2 とは異なり, モデル選択をしなくてもいいという利点がある. また, 元々上手く機能しているモデルや学習アルゴリズムに対して新しいパラメータの推定手法を考案するので, 新たに第 3 のアプローチを適用しようとした場合には, データの情報が要らないという利点がある.

ここで確率的な状態遷移と確率的な記号出力によってモデル化されている Hidden Markov models(HMM; 隠れマルコフモデル) を考える. HMM は数学的構造が明確であるため, 様々な事象に対して理論的な記述が可能である. そのため教師なし学習における代表的なモデルとして位置付けられている. HMM における学習は，データに対する尤度を最大化するパラメータを求める最適化問題となるが, 局所解が多く存在してしまうという問題がある. この問題に対し，従来は異なる初期状態から学習を開始し，複数の学習器を作る. これらの学習結果のうち, 尤度の最も高いパラメータを用いるという手法を行うが，このような方法では使わなかった学習器の学習が無駄になる. さらにデータに対して過学習を起こす可能性が存在する.

本論文では複数の学習器から，より過学習に強く，性能の高い HMM のパラメータ推定を行う第 3 のアプローチを提案する. そして状態ラベルの対応付けの問題の解決によって, 初めて具体的なアルゴリズムとして実現し, 効果を確認した.

2 提案手法の位置付け

本章では, 提案手法である隠れマルコフモデルにおける集合型パラメータ推定の位置付けを述べる.

アンサンブル学習としての位置付け
アンサンブル学習とは, 複数の結果がある時にそれらを用いてより良い結果を得る為の手法である. すなわち, 与えられたデータ集合 X に対する各々の予測 $p_i(X|\theta_i)(\theta_i$ はパラメータ; $i = 1, \cdots)$ を用いて, 効果的な新しい予測 $p_E(X)$ を得ることと定義できる. 代表的な例は $\sum_i \pi_i = 1$ という条件の下で, $p_E(X) = \sum_i \pi_i p_i(X|\theta_i)$ として新しい予測を構成する手法である. これに対し, 本論文で提案する手法は複数の推定結果であるパラメータ $\{\theta_i\}$ を用いて, より良いパラメータ θ_E を推定する手法である. すなわちアルゴリズム Λ によって, $\Lambda(\{\theta_i\}) = \theta_E$ として新たなパラメータを推定する. 複数のものを組み合わせ新しいものを生み出すという観点から見れば, 提案手法はアンサンブル学習の枠組みと見なすことができる.

正則化としての位置付け
正則化とは, 新しい情報 (例:罰金項) を用いることで, 過学習を防ぐ手法である. 代表的な例は Maximum a posteriori(MAP) 推定と呼ばれる. パラメータ θ に関して事前分布を導入することで, 最尤推定解 $\mathrm{argmax}_\theta \log p(X|\theta)$ を MAP 推定解 $\mathrm{argmax}_\theta\{\log p(X|\theta) + \log p(\theta)\}$ として θ を推定することができる. これに対し, 本論文で提案する手法は過学習を防ぐために最尤推定解を直接は採用しない. 具体的には平均を取ることでパラメータを一様に近づけている. すなわち複雑なモデルは正しくないという情報を用いることで, 新たなパラメータを推定している. 新たな情報を付け加えることで過学習を抑えるという観点から見れば, 提案手法は正則化の枠組みと見なすことができる.

3 隠れマルコフモデルにおける集合型パラメータ推定

本章では提案手法の具体的な手順を示す.

3.1 隠れマルコフモデル

有限の状態ラベル $\{1, \ldots, S\}$ と有限の単語集合 V が与えられた時, 隠れマルコフモデル (HMM) は状態遷移確率行列 $A(= a_{ij})$, 単語出力確率行列 $B(= b_{iv})$, 初期状態確率ベクトル $\pi(= \pi_i)$ で表される. ただし, a_{ij} は状態 i から j への遷移確率で,$\sum_j a_{ij} = 1$, b_{iv} は状態 i で単語 v を出力する確率であり,$\sum_v b_{iv} = 1$, π_i は状態 i が初期状態である確率で,$\sum_i \pi_i = 1$ である.

3.2 並列化

N 個の学習器を訓練データに対する尤度が大きい順に並べ替え, 大きいほうから順に $M^{(1)}, \cdots, M^{(N)}$ と呼ぶ. それぞれの学習器 $M^{(k)}(k = 1, \cdots, N)$ は状態遷移確率行列 $A^{(k)}(S \times S)$, 単語出力確率行列 $B^{(k)}(S \times V)$, 初期状態確率ベクトル $\pi^{(k)}(S$ 次元) をもつ. 単語出力確率行列を

$$B^{(k)} = \begin{bmatrix} b_{11}^{(k)} & \cdots & b_{1V}^{(k)} \\ \vdots & \ddots & \vdots \\ b_{S1}^{(k)} & \cdots & b_{SV}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1^{(k)} \\ \vdots \\ \mathbf{b}_S^{(k)} \end{bmatrix} \tag{1}$$

と表す. ただし, $b_{iv}^{(k)}$ は k 番目の学習器における, 状態 i で単語 v を出力する確率で, $\mathbf{b}_i^{(k)}$ は行ベクトルである.

3.3 類似度

学習器同士の状態ラベルを対応付けるために, 観測している単語情報から特徴を掴みやすい $\mathbf{b}_i^{(k)}$ を用いる. $\sum_i p_i = 1$, $\sum_i q_i = 1$ の 2 つのベクトル p, q が与えられた時, Hellinger 距離は $H(\mathbf{p}, \mathbf{q}) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$ と定義される. これを用いて, 類似度行列を $(h_{ij}^{(k)})$ と定義する. ただし, $h_{ij}^{(k)} = H(\mathbf{b}_i^{(1)}, \mathbf{b}_j^{(k)})(i, j = 1, \cdots, S)$ である.

3.4 提案手法

この節では類似度行列を用いた, 具体的な状態ラベルの対応付けについて示す. 対応付けの方法の違いによって 3 つの手法を提案する. 対応を表すため, $M^{(1)}$ の状態ラベル $i(i = 1, \cdots, S)$ と $M^{(k)}(k = 1, \cdots, N)$ の状態ラベル $j(j = 1, \cdots, S)$ が対応する時, $q_{ij}^{(k)} = 1$ そ

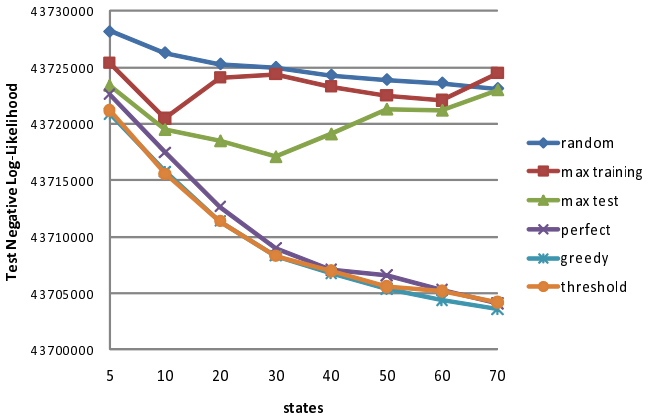

図 1: 状態数による Negative Log-Likelihood.

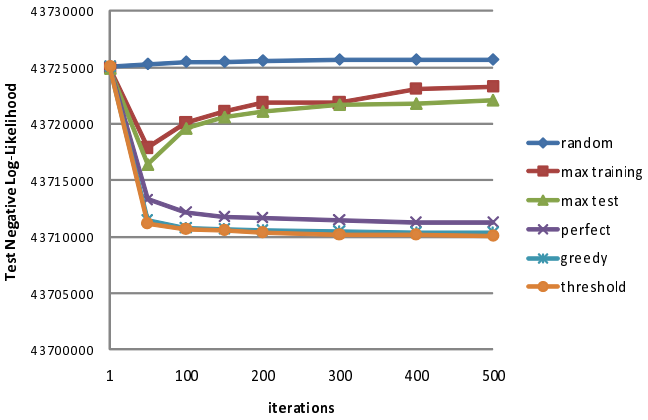

図 2: 繰り返し回数による Negative Log-Likelihood.

れ以外の時 $q_{ij}^{(k)} = 0$ となる類似度指標 $\{q_{ij}^{(k)}\}$ を定義する.

完全マッチング対応付け手法

1 つ目の提案手法では, 類似度行列を用いた完全マッチングを行う. 類似度の総和が最小となるように, 学習器同士のそれぞれの状態を 1 対 1 に対応させる. これを完全マッチング対応付け手法と呼ぶ. 最適化問題として定式化すると

$$\left\| \begin{array}{l} \arg\min_{q_{ij}^{(k)}} \sum_{i=1}^{S}\sum_{j=1}^{S} q_{ij}^{(k)} h_{ij}^{(k)} \\ \text{subject to} \quad \sum_{j=1}^{S} q_{ij}^{(k)} = 1 \quad (i = 1, \cdots, S) \\ \qquad\qquad \sum_{i=1}^{S} q_{ij}^{(k)} = 1 \quad (j = 1, \cdots, S) \\ \qquad\qquad q_{ij}^{(k)} \in \{0, 1\} \end{array} \right. \tag{2}$$

この問題は完全二部グラフで両側の頂点数が同じ場合の最小重みマッチングとなるのでハンガリアン法 [2] で解いて, 解 $\{q_{ij}^{(k)}\}$: を求める.

緩い対応付け手法

2 つ目の提案手法では, 類似度行列の各行から 1 つ, 最小の値を選ぶ対応付けを行う. これを緩い対応付け手法と呼ぶ. 最適化問題として定式化すると

$$\left\| \begin{array}{l} \arg\min_{q_{ij}^{(k)}} \sum_{i=1}^{S}\sum_{j=1}^{S} q_{ij}^{(k)} h_{ij}^{(k)} \\ \text{subject to} \quad \sum_{j=1}^{S} q_{ij}^{(k)} = 1 \quad (i = 1, 2, \cdots, S) \\ \qquad\qquad q_{ij}^{(k)} \in \{0, 1\} \end{array} \right. \tag{3}$$

行ごとに距離に応じた値を選択すると解は, $\{q_{ij}^{(k)}\} : q_{ij}^{(k)} = 1\big(j = \arg\min_{u=1,\cdots,S} h_{iu}^{(k)},\ q_{ij}^{(k)} = 0(\text{otherwise})\big)$ となる.

閾値対応付け手法

3 つ目の提案手法では, ベクトル間の Hellinger 距離に対してある閾値 τ を与え, その値未満となった状態ラベル全てを対応付ける. これを閾値対応付け手法と呼ぶ. 解は以下のようになる.

$$q_{ij}^{(k)} = \begin{cases} 1 & \text{if } h_{ij}^{(k)} < \tau \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

3.5 平均化

$\{q_{ij}^{(k)}\}$ に基づき, 新しい学習器を作る. 完全マッチング対応付け手法では全てのパラメータに関して平均したパラメータを用いる. また, 緩い対応付け手法, 閾値対応付け手法では, 重複を許した対応付けであるため単語出力確率行列, 初期状態確率ベクトルは平均したパラメータを用い, 状態遷移確率行列は $M^{(1)}$ のパラメータを用いる.

4 実験

データセットとして Wall Street Journal(11000 sentences[training 10000, test 1000]) を用いる. 評価はテストデータに対する negative log-likelihood で行った. Hellinger 距離の結果のみ図で表した.

状態数による評価

まず, 状態数の増加による negative log-likelihood の変化について評価した. EM アルゴリズムにおける繰り返し回数は 100, 並列数は 20 と固定した. 追実験では状態数を追加した. 図 1 はどの提案手法も従来手法に比べ, 尤度の値が良くなったことを示している. また, 状態数が増加するにつれて, negative log-likelihood の下がり方が増加する傾向にあることも示している.

繰り返し回数による評価

次に, EM アルゴリズムにおける繰り返し回数の増加による negative log-likelihood の変化について評価した. 状態数は 20 と固定した. 追実験では並列数を 20 とし, 状態数に関する実験と条件を揃えた. 図 2 は繰り返し回数が増加するに連れて, 従来手法が単一の学習器であるために, 過学習を起こすことを示している. 逆に対応付けを行い, 集合型パラメータ推定を行うことで, パラメータは一様に近づき, 過学習を抑えていることも示している.

5 おわりに

本論文では異なる初期値から学習した複数の学習器のパラメータを平均することで, より過学習に強く, 性能の良いパラメータを新たに推定する手法を HMM に適用した. 具体的には状態ラベルの対応付けの違いによって, 完全マッチング対応付け手法と緩い対応付け手法, 閾値対応付け手法を提案し, これに従来手法を加えて, 評価実験により比較した. 実験の結果, いずれの提案手法も従来手法より汎化性能の良いパラメータが得られることが分かった.

参考文献

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] F. Bourgeois and J.-C. Lassalle. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, Vol. 14, No. 12, pp. 802–804, 1971.