

音響モデルのアンサンブル学習

篠崎 隆 宏^{†1}

機械学習において広く用いられているアンサンブル手法を隠れマルコフモデルを用いた音響モデルの構造最適化やパラメータ推定に応用する研究について紹介する。

Acoustic Model Training based on Ensemble Methods

TAKAHIRO SHINOZAKI^{†1}

This paper introduces a series of studies that apply ensemble methods in parameter and structure optimizations of hidden Markov models.

1. はじめに

機械学習におけるアンサンブル法は、複数の識別器を用いることで単独の識別器を用いたときよりも高い予測精度を得ようとする手法である。複数の識別器は学習セットをサブサンプリングしたり、入力特徴量として異なるものを用いたり、パラメータ数や学習初期値を変えて異なるモデルを学習したりすることにより作成される¹⁾。代表的な手法としてはバギング²⁾やブースティング³⁾があげられる。バギングでは複数の識別器を学習セットからランダム抽出したサブセットより学習し、評価時にはそれら複数の識別器による結果を投票やスコアの和をとることで統合し最終的な出力を得る。またブースティングの代表例であるアダブースト⁴⁾では逐次的に学習サンプルの重みを変えながら異なる識別器を学習し、評価時にはそれらを重みづけして組み合わせて用いる。モデル選択法として用いられるクロスバリデーション法 (CV 法) もアンサンブル法の一つであり⁵⁾、モデル集合中で一つのモデルの

み重み 1 それ以外は重み 0 でモデル統合を行っているときもみ出すことも出来る。

音響モデルにおける典型的なアンサンブル手法の利用の試みとしては発話レベルでアダブーストを適用した研究が挙げられ、大語彙音声認識により有効性が示されている⁶⁾。他方、CV 法による隠れマルコフモデル (HMM) の構造決定については、これまで計算量の問題などから連続分布 HMM を対象とした大規模な応用は限られていた。また、通常アンサンブル法は評価時の性能を直接的に向上させるために用いられるが、EM 法などの確率評価を内包する繰り返し学習アルゴリズムの内部に応用することで学習の汎化性能の向上を図り、認識時にはそのようにして学習された単独のモデルを使用することも考えられる。この場合、認識時におけるモデルの扱いに変更がなく従来のデコーダがそのまま使用出来、また複数のモデルを用いることによる認識時の計算コストの増大を避けられる利点がある。以下では CV 法によるモデル選択を発展させ混合ガウス分布 HMM の状態数や混合数を効率的に最適化する手法や、アンサンブル的な手法を繰り返し最尤推定法の内部に組み込むことでモデルの推定精度を向上させる手法に関して、これまでに行った研究を紹介する。

2. アンサンブル法による HMM の構造最適化

HMM を用いた音響モデルにおいて、与えられた学習データを有効に活用するためにはモデルの状態数や混合数を学習データに対して最適化することが重要である。HMM 状態数の最適化法としては状態集合を順次分割していく決定木クラスタリング⁷⁾が広く用いられている。分割基準としては学習セットに対する尤度が用いられることが多いが、学習セットに対する自己尤度は状態集合の分割に対して単調増加であり、それ自体では適切な状態数を決定することが出来ない問題がある。停止基準を与えるために情報量基準を用いることも考えられるが⁸⁾、本来情報量基準による手法は混合ガウス分布や HMM のような非正則なモデルは適用対象外であり、最適性は保証されない。即ちモデルが正則である場合には学習セット $X = \{x_1, x_2, \dots, x_N\}$ から最尤推定により求めたパラメータ集合が $\hat{\theta}$ であるとするとき、式 (1) に示すように $\hat{\theta}$ の真のデータ分布に関する期待対数尤度の不偏推定量を学習セットに対する尤度とパラメータ数 k から近似的に求めることが出来るが⁹⁾、モデルの正則性が仮定できない場合にはそのような近似は成り立たない。

$$E_X \left[\left(\frac{1}{N} \sum_i \log (P(x_i | \hat{\theta})) - \frac{k}{N} \right) - E_y [\log (P(y | \hat{\theta}))] \right] \approx 0. \quad (1)$$

これに対し学習データの使用方法の工夫に基づく Leave-one-out CV 法を用いた場合は

^{†1} 東京工業大学
Tokyo Institute of Technology

式 (2) に示すように、ほぼ（サンプル数が 1 だけ少ない点を除いて）完全にバイアスのない推定値を得ることが出来る。

$$E_X \left[\frac{1}{N} \sum_i \log (P(x_i | \hat{\theta}_i)) - E_y [\log (P(y | \hat{\theta}))] \right] = 0. \quad (2)$$

ここで $\hat{\theta}_i$ は x_i を除いて求めたモデル、 $\hat{\theta} (\approx \hat{\theta})$ は 1 サンプルだけ少ない学習セットから求めたモデルである。

HMM の状態最適化においては非常に多くのモデルが比較対象となるために、それら全てのモデルを逐一学習セットから学習することは非現実的である。従来の自己尤度による決定木クラスタリングは、ガウス分布の十分統計量を用いることで高速に実行される。これはアライメント固定の仮定の元で、決定木クラスタリングによりクラスタリングされた状態を持つ任意の HMM のパラメタおよびその学習セットに対する尤度が、ともに事前に計算された十分統計量のみを用いて計算できることに基づいている。文献 10) では十分統計量を用いた高速な計算方法が CV 尤度の計算にも拡張可能であることを示し、また大語彙音声認識実験により CV 尤度を用いることで状態数を自動的に決定しつつ従来の尤度を用いた場合よりも高い認識性能が得られることを示した。同様のアルゴリズムは混合ガウス分布の最適化にも用いることが出来、文献 11) では HMM の各状態ごとに最適化を行う手法を提案し、認識実験において有効性を示した。しかし一方で、非常に多数のモデルの比較に CV 法を用いることの問題点も明らかになった。即ち CV による各モデルの尤度評価値は不偏であるが分散を持つ為、多数の CV 尤度の最大値には正の偏りが生じ、結果として過学習の問題が生じる。実際、混合要素対の選択併合に基づく最適化では m 個の要素を $m-1$ 個に減らす過程で $\frac{m(m-1)}{2}$ の多数の混合モデルが比較される。そこで尤度評価値の分散を減少させることでより最適化の性能を向上させることを目的として、文献 12) では CV にバギングのアイデアを導入した混合ガウス分布の最適化手法の提案を行い、自己尤度や MDL 基準を用いた場合と比較して高い認識性能が得られることを示した。

3. アンサンブル法による HMM の教師あり学習

HMM のパラメタ推定には繰り返し最尤推定法である EM アルゴリズム¹³⁾ が用いられる。HMM の推定においては、EM アルゴリズムは適当な初期モデルの元でガウス分布や状態遷移確率分布に関する十分統計量の期待値を計算するステップ（E ステップ）と得られた期待十分統計量から最尤推定モデルパラメタを求めるステップ（M ステップ）を交互に繰

返す操作となる。EM アルゴリズムでは学習セットに対する尤度が学習の繰り返しとともに単調増加することが保証されている。しかし、学習セットに対する尤度は式 (3) に示すように真の分布に対する尤度とバイアス項 α の和であり、真の分布に対する尤度を最大化するモデルが得られるわけではない。

$$\frac{1}{N} \sum_i \log (P(x_i | \theta)) = E_y [\log (P(y | \theta))] + \alpha(X, \theta). \quad (3)$$

EM アルゴリズムのプロセスを観察すると、E ステップで使用されるモデルはその前の M ステップで同じ学習データから学習されたモデルであるため尤度評価におけるバイアスの原因となり、かつその影響は学習の繰り返しとともに増大してしまう問題があると考えられる。そこで、E ステップにおける尤度評価の精度を向上させ学習汎化性能を向上させる目的で、CV 法およびバギングに似た手法をそれぞれ EM の枠組内に導入した CV-EM 法および Ag-EM 法を文献 14) において提案した。CV-EM は E ステップと M ステップにおいて使用するデータを分離することで、Ag-EM は E ステップにおいて複数のモデルによる尤度評価を統合することで、学習汎化性能の向上を図る。提案法は EM 法と比べてより精密なモデルをより精度良く推定することが出来、認識性能の向上に有効であることを示した。

4. アンサンブル法による HMM の教師なし学習

音響モデルの教師無し話者適応の枠組として、初期モデルによる認識結果をラベルとして MLLR¹⁵⁾ などの教師あり適応を行い、さらに得られた適応モデルを元に認識と教師あり適応を繰り返すバッチ型適応が広く用いられている。このようなバッチ型教師無し適応はパラメタ拘束やビタビ近似を取り入れた EM アルゴリズムととらえることが出来、CV-EM や Ag-EM を応用することが可能である。そのような方法として、文献 16) では教師無し CV 適応法および教師無し集合適応法の提案を行った。提案法は従来のバッチ型適応で問題となる、認識仮説に含まれる認識誤りを含めて推定されたモデルが次の認識ステップで同じ認識誤りを生じ、さらにその影響が適応ループ中で繰り返し強化されてしまう問題を効果的に抑制することが出来る。提案法を教師無し話者適応やタスク適応に応用し、従来のバッチ型適応と比べて認識性能向上に大きな効果があることを示した。

5. ま と め

アンサンブル手法を隠れマルコフモデルを用いた音響モデルの構造最適化やパラメタ推定に応用する研究について紹介を行った。

参 考 文 献

- 1) T.G. Dietterich. *Ensemble Methods in Machine Learning*, volume 1857/2000. Springer Berlin / Heidelberg, 2008.
- 2) L.Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- 3) RobertE. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
- 4) Y.Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- 5) Machine learning ensemble. http://en.wikipedia.org/wiki/Machine_learning_ensemble, June 2009.
- 6) C.Meyer and H.Schramm. Boosting acoustic models in large vocabulary speech recognition. *Speech Communication*, 48(5):532–548, 2006.
- 7) S.Young, J.Odell, and P.Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Workshop on Human Language Technology*, pages 307–312, 1994.
- 8) K.Shinoda and T.Watanabe. Acoustic modeling based on the MDL principle for speech recognition. In *Proc. EuroSpeech*, volume1, pages 99–102, 1997.
- 9) H.Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- 10) T.Shinozaki. HMM state clustering based on efficient cross-validation. In *Proc. ICASSP*, volumeI, pages 1157–1160, Toulouse, 2006.
- 11) T.Shinozaki and T.Kawahara. Gaussian mixture optimization for HMM based on efficient cross-validation. In *Proc. Interspeech*, pages 2061–2064, 2007.
- 12) T.Shinozaki, S.Furui, and T.Kawahara. Aggregated cross-validation and its efficient application to Gaussian mixture optimization. In *Proc. Interspeech*, pages 2382–2385, 2008.
- 13) A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society*, (Series B 39, No. 1):1–38, 1977.
- 14) T.Shinozaki and M.Ostendorf. Cross-validation and aggregated EM training for robust parameter estimation. *Computer speech and language*, 22(2):185–195, 2008.
- 15) C.J. Leggetter and P.C. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. Eurospeech*, pages 1155–1158, 1995.
- 16) 篠崎 隆宏, 久保田 雄, and 古井 貞熙. 高精度音声認識のための教師なしクロスバリデーションおよび集合適応法の提案. In **情報処理学会 研究報告**, volume 2009-SLP-75, pages pp.1–6, 2 2009.