

平成 21 年 5 月 29 日現在

研究種目：若手研究 (B)  
 研究期間：2007～2008  
 課題番号：19700167  
 研究課題名 (和文) クロスバリデーション尤度を用いた統計的パターン分類器学習アルゴリズムの研究  
 研究課題名 (英文) Statistical pattern classifier training based on cross-validation likelihood  
 研究代表者  
 篠崎 隆宏 (SHINOZAKI TAKAHIRO)  
 東京工業大学・大学院情報理工学研究科・助教  
 研究者番号：80447903

## 研究成果の概要：

モデル選択手法として用いられているクロスバリデーション的手法を繰り返し最尤パラメタ推定法の内部に効果的に組み込むことで、少ないデータから精密かつ一般性の高い統計モデルを学習するための手法を提案した。具体的には期待値最大化学習法にクロスバリデーションを組み込んだ手法の提案を行い、さらに教師なし適応への応用も行った。提案手法を音声認識で用いる統計モデルの学習に応用し、認識性能の向上に非常に効果的であることを示した。

## 交付額

(金額単位：円)

	直接経費	間接経費	合計
2007年度	1,700,000	0	1,700,000
2008年度	1,600,000	480,000	2,080,000
年度			
年度			
年度			
総計	3,300,000	480,000	3,780,000

## 研究分野：総合領域

科研費の分科・細目：情報学 知覚情報処理・知能ロボティクス

キーワード：統計モデル、過学習、汎化性能、クロスバリデーション、隠れマルコフモデル、期待値最大化学習

## 1. 研究開始当初の背景

統計的手法を用いたパターン認識は、現実世界の不確かな現象を計算機で取り扱う有効な方法として音声認識や画像処理、遺伝子解析など様々な分野で用いられている。統計的モデルは尤度最大化基準などにより学習データより作成されるが、この際に一般的に問題となるのがモデル構造やパラメタが学習データに過度に特化してしまう過学習である。

従来この問題に対する対策として、モデルのパラメタ数を情報量基準などにより制御することが行われてきた。しかし、幾つかの理論的仮定に基づく情報量基準は実際の問題への応用において、しばしば理論値からのずれを補う経験的な補正項を必要とし、さらには幅広い応用を持つ隠れマルコフモデル(HMM)などの隠れ変数を含んだモデルの比較には本質的に役に立たないという欠点がある。これは有限のパラメタ数に対しても尤度が無限大に発散してしまう不安定性の間

題が存在するためである。

これに対しデータドリブンな方法であるクロスバリデーションは様々な問題に原理的に適応可能である長所を持つ反面、その利用は複数の条件で作成したモデルを比較するなど、学習アルゴリズムの後处理的な応用などに限られていた。これは大規模なデータを処理する上で必須となる従来の学習アルゴリズムの基本的な高速化手法が、クロスバリデーションを導入することによりしばしば使用できなくなってしまう、計算量が非現実的に大きくなってしまいうためである。また、そもそもクロスバリデーションをモデルの学習アルゴリズム内部に組み込もうとする試みは、限定的な条件を対象とした数例を除きほとんど行われてこなかった。

## 2. 研究の目的

本研究では、従来の自己尤度（学習セット全体から学習されたモデルの、学習セット全体に対する尤度）最大化基準に代えて、クロスバリデーション尤度最大化基準を用いてモデル推定を行うことで、モデルサイズの決定のみならず、不安定性を含め過学習の問題一般を解消し、モデル性能を向上させることを目的とする。具体的には、

- (1) 十分統計量を活用することで期待値最大化学習法の内部に効率的にクロスバリデーションを導入するアルゴリズムの提案し、またその改良を行う。
- (2) 提案アルゴリズムを実装したプログラムを作成する。
- (3) 提案アルゴリズムを音声認識システムにおいて使用される統計的音響モデル（隠れマルコフモデル）の学習に応用することで、認識性能の向上を目指す。

## 3. 研究の方法

- (1) アルゴリズムについて、理論的、実験的な検討を行うことで、問題点を洗い出し、さらなる改良を目指す。
- (2) 提案アルゴリズムを実装するプログラムは大規模な統計モデルの学習を可能とするため、クラスターコンピュータ（並列式スーパーコンピュータ）を対象とした実装とする。ソフトウェア開発を効率的に行うため、既存のツール

キット等を極力活用する。とくに、本研究で検証実験の対象とする隠れマルコフモデルの学習において、提案法は自己尤度を用いた従来の学習プロセスとプログラム上共通した要素が多く、既存のソフトウェアを活用することでソフトウェアの開発プロセスの大幅な効率化が可能である。

(3) 提案アルゴリズムの評価実験は、提案アルゴリズムを音声認識システムにおいて用いられる隠れマルコフモデルを用いた音響モデルの学習に応用することで行う。また、音声認識のタスクは、現在音声認識分野において活発に開発競争が行われている自然発話音声を対象とした大語彙連続音声認識とする。このためには、音響モデルの学習には数百時間の音声データが必要となり、モデル学習用のデータとしては、日本語における世界最大のデータベースである「日本語話し言葉コーパス」を用いる。従来の学習法、および提案法を、それぞれの方法で学習した音響モデルを用いた認識システムの認識性能を比較することで評価する。

## 4. 研究成果

クロスバリデーションを用いた提案学習法およびそれを発展させたアルゴリズムとして、クロスバリデーション尤度を用いた効率的な混合ガウス分布のモデル構造最適化法、クロスバリデーションおよびバギングと似た手法をそれぞれ期待値最大化学習法の枠組み内に組み込んだクロスバリデーションEM法および集合EM法、クロスバリデーションEM法を教師なし学習法に応用した教師なしクロスバリデーション適応法の提案を行い、音声認識実験により評価を行った。

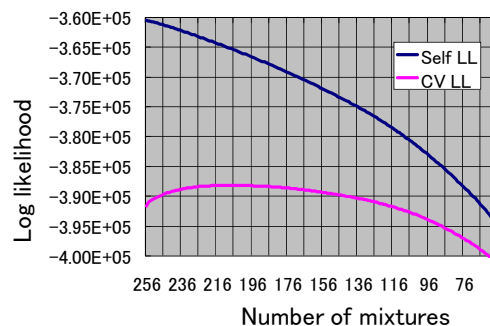


図 1

図 1 に混合ガウス分布の過剰な混合要素

の削減に従来の自己尤度を用いた場合と提案するクロスバリデーション尤度を用いた場合の尤度変化の様子を示す。従来法はモデルの学習に使用したデータを用いて評価尺度を計算するため、いわば答えを見てから試験を受けるようなものであり、得られる評価スコアは実際のモデルの性能よりも大きなものとなる傾向がある。また同様の理由でパラメタ数の多いモデルほど高いスコアを与える傾向があり、従来法では経験的な閾値操作などと組み合わせなければ最適な混合要素数を決定できない問題がある。これに対し、提案法を用いた場合学習データのほぼ全体をモデル作成およびその評価に活用しつつ、評価データと学習データを分離することができ、モデルの性能を反映した信頼性の高い評価スコアを求めることができる。これは一般のクロスバリデーション法の利用と同じ利点であり、本研究で提案した十分統計量を用いた効率的なアルゴリズムにより、初めて大規模な混合ガウス分布の最適化に適用可能となった。クロスバリデーションにより求めたスコアは自己尤度法と異なり混合モデルの要素数に対して最大値を持つ。スコアの最大値はその点でモデルの性能が最大になることを示しており、クロスバリデーションスコアを用いることで、よりよい混合要素の配置と混合要素数を自動的に決定することができる。

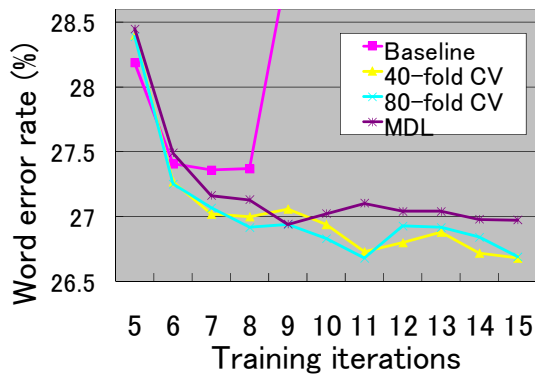


図 2

図 2 に混合ガウス分布の最適化法を従来の自己尤度を用いて行った場合と提案法を用いて行った場合の単語誤り率 (Word error rate) の比較を示す。単語誤り率が低いほど、認識性能が高い。横軸は学習のステージ数で、混合ガウス分布の最適化は学習ステージごとに行っている。また、混合ガウス分布の過剰な要素を削減する最適化処理を行わない場合 (図で baseline) は、学習のステージごとにモデルの混合数が 2 倍となる。この場合、モデルのパラメタ数 (混合数) が少ないうち

は学習ステージの増加とともにモデルの性能が向上するが、あるところまで増えると、逆に学習ステージとともにモデルの性能が低下するのが図より分かる。これはモデルのパラメタ数が大きいほどより精密なモデル化が可能になるが、一方で多数のパラメタを正確に推定するためには大量のデータが必要であり、学習データ量が一定の条件ではモデルのパラメタ数が大きくなりすぎるとそれらを正しく推定できなくなってしまうためである。したがって、与えられたデータ量に応じてモデルのパラメタ数を最適化することが非常に重要となる。提案法を用いることで混合ガウス分布のモデルパラメタ数は自動的に制御され、また各要素分布の配置も効果的なものとなる。これにより、提案法を用いることで従来よりも高い認識性能を学習ステージ数に対して頑健に得られるようになった。また提案法ではクロスバリデーションの回数 (図では 40 と 80) の選択についても、ある程度大きく取れば安定した性能が得られることが示された。さらに情報量理論に基づく別のアプローチである MDL 基準法と比較しても、提案法の方が高い認識性能が得られることが示された。

図 3 に期待値最大化アルゴリズム内にクロスバリデーション法を組み込んだクロスバリデーション EM 法の学習プロセスを示す。図において「M」はモデルを表し、「SS」は十分統計量を表す。提案手法では学習データを K 個の互いに排他的な区画に分割し、区画ごとに統計量を計算する。これら区画ごとに求めた統計量をもとに、学習の繰り返しにおいて期待値ステップと最尤ステップでデータの重複を避けることで、学習プロセスの汎化性能を高める。

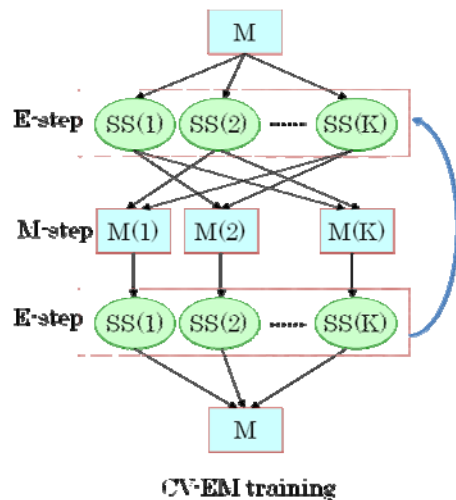


図 3

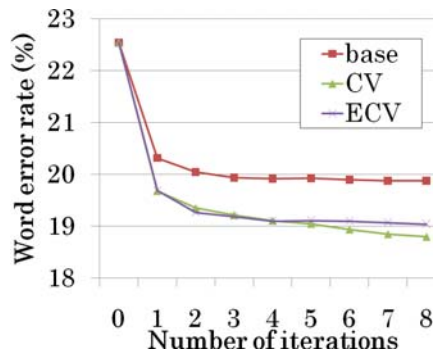


図 4

図 4 に音声認識における教師なし話者適応に、提案法である教師なしクロスバリデーション適応法を応用した結果を示す。図で「base」は従来の繰り返し法に基づく話者適応の結果、「CV」がクロスバリデーション法を応用した提案法による結果、および「ECV」が CV 適応法において若干の近似を導入することで計算量を大幅に削減する高速化版アルゴリズムを用いた場合の結果である。どの適応においても、モデルパラメタの更新は最尤線形回帰法に基づいており、違いは教師なし適応のループ中でのデータの扱いである。クロスバリデーション適応法では、繰り返しのループ中で行われるモデル推定と認識処理でデータを分離することで、従来法よりも高い話者適応効果を得ることができ、認識性能が向上していることが分かる。

#### 5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文] (計 6 件)

- ① Takahiro Shinozaki, Yu Kubota, and Sadaoki Furui, "Unsupervised Cross-validation Adaptation Algorithms for Improved Adaptation Performance," Proc. ICASSP 2009, pp. 4377-4380, (2009).  
査読あり
- ② 篠崎 隆宏, 久保田 雄, 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」, 情処研報, 2009-SLP-75, pp. 1-6 (2009).  
査読なし

- ③ Takahiro Shinozaki, Sadaoki Furui, and Tatsuya Kawahara, "Aggregated Cross-validation and Its Efficient Application to Gaussian Mixture Optimization," Proc. Interspeech 2008, pp. 2382-2385 (2008).  
査読あり

- ④ 篠崎 隆宏, 古井 貞熙, 河原 達也, 「効率的なクロスバリデーションに基づく混合ガウス分布の最適化法とその拡張」, 情処研報, 2008-SLP-72, pp. 69-74 (2008).  
査読なし

- ⑤ 篠崎 隆宏, 河原達也, 「効率的なクロスバリデーション尤度評価に基づく混合ガウス分布の最適化」, 情処研報, 2007-SLP-67, pp. 81-86 (2007)  
査読なし

- ⑥ 篠崎 隆宏, Ostendorf, Mari, 河原 達也, 「頑健なパラメタ推定のための Aggregated EM 法の提案と評価」, 信学技報, Vol. 107, pp. 223-228 (2007)  
査読なし

[学会発表] (計 2 件)

- ① 篠崎 隆宏, 久保田 雄, 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適応法の提案」, 日本音響学会, 2009年3月17日, 東京工業大学
- ② 篠崎隆宏, 古井貞熙, 河原達也, 「Aggregated cross-validation 尤度を用いた混合ガウス分布最適化アルゴリズムの提案」, 日本音響学会, 2008年3月18日, 千葉工業大学

[その他]

ホームページ

<http://www.furui.cs.titech.ac.jp/~shino t>

#### 6. 研究組織

(1) 研究代表者

篠崎 隆宏 (SHINOZAKI TAKAHIRO)

東京工業大学・大学院情報理工学研究所・助教

研究者番号: 80447903

(2)研究分担者  
なし

(3)連携研究者  
なし