論文の内容の要旨

論文題目　有限混合分布モデルの学習に関する研究

氏名　赤穂 昭太郎

社会の情報化に伴い，膨大かつ複雑なデータから有用な情報を抽出するための知的な情報処理手法が求められている．そのような状況の中で注目されているのがグラフィカルモデルと呼ばれる統計モデルであり，統計学や統計物理学，あるいは脳科学や人工知能といった分野を中心とした学際領域で盛んに研究されている．本論文で扱う (有限) 混合分布モデルは最も単純なグラフィカルモデルの一種である．その形の単純さに反し，より複雑なグラフィカルモデルの基本的な性質を受け継いでおり，理論的な解析などにおいて重要なモデルと位置付けられている．混合分布は，複数の確率分布の重ね合わせで定義される確率モデルで，複雑な分布を単純な分布の和に分解して記述するという分割統治的な性質を持つ．本論文では混合分布モデルの学習 (特に最尤推定) に関して，汎化と学習アルゴリズムという理論的な側面に関する解析と，パターン的な情報に埋もれたシンボル情報を復元するという混合分布モデルの応用的な側面についての検討を行った．

まず最初に，あるクラスの混合分布の持つ汎化能力が従来考えられてきたのとは異なる振舞いを示すことを明らかにした．汎化能力とは，モデルが有限個の訓練サンプルにフィットするだけでなく，背後にある真の分布をどれだけ表現できるか

という能力である．統計モデルの汎化能力の評価は，例題からの学習に起因する基本的な問題であり，統計学や情報理論などの枠組みからさまざまな研究がなされてきた．特に最近ではニューラルネットモデルや混合分布など特異性を持つモデルに関する汎化能力の研究に注目が集まっている．一般に，統計モデルの構造を複雑にすればするほど訓練データに対する当てはまり具合はよくなる．しかし逆に，過度に適合した統計モデルでは汎化能力は低くなってしまう．したがって，訓練サンプルに適合するだけではなく，できるだけ「単純な」モデルを選ぶことによって，汎化能力を高く保つ必要があると考えられている．ところが，本論文では，Radial Basis Boltzmann Machine (RBBM) と呼ばれる特殊な形の正規混合分布では，これに反する現象が観察されることがわかった．RBBM では，要素分布である正規分布の分散がモデルの複雑度を制御するコントロールパラメータになっており，それを変化させることにより分岐現象を起こし，モデルパラメータの見かけの個数を調節することができる．本論文ではまず，分岐の振舞いを解析し，適当な仮定のもとで 4 次のキュムラントに依存して分岐の様相が特徴づけられることを示した．次にその結果に基づいて，真の尤度と経験尤度とのバイアスを表す竹内の情報量規準 (TIC) を用いて，RBBM の汎化能力を評価した．すると，分岐する前では直観的に予測されるようにモデルの複雑度が増えるにつれてバイアスも増加するという性質が導かれたが，分岐した直後では分岐によって見かけのパラメータ数が増加するにも関わらず $-\infty$ の傾きでバイアスが減少する場合が存在することが解析的に示された．すなわち，この場合には分岐した後のより「複雑な」モデルの方が分岐する前の単純なモデルよりも汎化能力が高いことを意味している．本論文では，こうして得られた結果を計算機実験によって確認した．

次に，混合分布の学習アルゴリズムとして知られる EM (Expectation-Maximization) アルゴリズムを，複雑な推定問題を単純にする手法であるととらえ，確率分布のパラメータ推定問題に応用した．EM アルゴリズムは，一般に欠測データがある場合などに最尤推定の局所最適解を得るための安定なアルゴリズムとして知られているが，最近になって，曲がった空間での最適化問題をより平坦な空間の最適化問題の繰り返し計算に還元させているという，幾何学的な意味も明らかにされてきた．実際，混合分布の EM アルゴリズムは，構成する個々の要素分布の推定問題というより簡単な問題に帰着される．ただし，その要素分布が正規分布のような単純な

分布では推定が簡単にできるが，そうでない場合はやはり難しい問題が残る．これを解決するためのアプローチとして，要素分布を再び混合分布でモデル化するという再帰的な手段を用いたのが Jordan らの階層的エキスパート混合モデルである．しかしながら，このアプローチでは一般にパラメータ数が多くなるため汎化の点で不利であるという点と，既に要素分布として基本となる分布形が与えられている場合には用いることができない点で問題がある．そこで本論文では，既に要素分布の形が与えられていた場合に EM アルゴリズムを応用した学習アルゴリズムが適用できないかと考え，任意に与えられた分布が位置・尺度等のパラメータを持つとき，その推定に EM アルゴリズムを適用する問題を調べた．その結果，任意の次元で位置・尺度パラメータを持つ場合と，2 次元で位置・尺度・回転パラメータを持つ場合という二つの場合について，閉じた形のアルゴリズムが得られた．まず，与えられた分布を必要な精度で，ある決められたクラスの正規混合分布で近似しておく．その上で EM アルゴリズムを適用するのだが，単純な正規混合分布のパラメータ推定と異なり，推定すべきパラメータが正規混合分布のすべての要素分布に含まれており，そのままでは閉じた形のアルゴリズムが得られない．そこで，EM アルゴリズムの拡張である ECM (Expectation-Constrained Maximization) アルゴリズムを適用すると，計算可能な統計量を係数として持つ 2 次方程式の解としてパラメータの推定更新式が得られることがわかった．これは，単純な正規混合分布の EM アルゴリズムが 1 次方程式の解として得られる場合の拡張ととらえることもできる．本論文では，こうして得られたアルゴリズムを，人工データおよび実画像を用いたテンプレートマッチングに応用し，有効性を確認した．

さて，本論文では最後に，混合分布を，パターン的な情報の中に隠れたシンボル的情報を扱うためのモデルととらえ，属性概念獲得課題を設定し，モデルの適用を試みた．実世界の情報が持っている大きな問題点の一つはその情報が不完全であるということである．特に，データを構造化するのに有用な記号的な情報は，生データに埋もれていて陽には得られない場合が多い．そのような状況下での学習の例題として，複数情報源からの属性概念獲得の問題を設定し，混合分布と EM アルゴリズムによる学習を実データに対して実験を行った．この問題では，音声と画像のペアのデータから，そこに隠された属性概念という構造を統計モデルを用いて学習する．具体的には，各画像に対しその属性 (色，形，大きさ等) の一つをランダ

ムに選んで発話した音声をペアとして与え (ただしどの属性であるかは与えない), そのような学習サンプルを数多く学習することにより属性概念を自立的に獲得するという課題である. この問題自体は発達心理学における概念形成のモデルと関連が深いが, 工学的には近年盛んに研究されている電子秘書などのアプリケーションにおける環境情報の学習の基礎となる研究である. また, ここで設定した問題はモダリティ間の情報を統合することによってはじめて可能となるという点で従来の単純な概念獲得課題よりも難しく新規性があるものである. 本論文では, まず各情報源毎に固有の特徴抽出を行い, 情報源間に共通して含まれる低次元の情報を抽出するための手段として正準相関分析を行った. その上で, 画像特徴から音声特徴への1対多写像を線形回帰モデルの混合分布でモデル化し, 学習を行った. 学習は EM アルゴリズムで行ったが, 不適切な局所最適解への収束を避けるため, この課題に特有の 1 対多写像であるという性質を利用した初期解の取り方を工夫した. その結果, 比較的基本的な手法の組合せであるにも関わらず, 未知サンプルに対しても高い学習性能結果が得られた.