4R-04

多次元正規分布のためのEM法

五十嵐　有真†　　三浦 孝夫†

法政大学理工学部創生科学科

東京都小金井市梶野町 3-7-2

1. 前書き

近年、インターネットの普及、観測技術の発達に伴い、数多くの経路から複雑で膨大なデータが生成されている。その結果,膨大な情報量となり,解析に必要な処理がより複雑かつ膨大という問題がある本研究では、多くの変量を持つ事象の解析における処理精度を向上するため、任意の多次元に拡張した正規分布により、モンテカルロ法を用いたEMアルゴリズム(以下EM法)を提案する。多次元正規分布を用いたサンプリングアプローチを行うことで近似的にデータの分布を求め、不完全な場合の推定を行う。またその有効性を処理精度により評価する。多次元正規分布を使う理由は次の2点ある。1つは全ての現象の基準が正規分布であること。もう1つは変量間の関連をモデルに取り込んでいることである。従来の EM 法では期待値計算を行う際に、膨大な計算を必要とし結果、計算効率を低下させる原因になる。本提案のアイデアは、分散共分散行列を利用したサンプリングを行うことで、変数間の相関を反映した値を得ることができることである。

2. 混合モデルと EM 法

混合モデルとは、複数の分布を重ね合わせることで作られる分布であり、複雑な構造のため単純に最尤推定を行ってもモデルの効果的な推定をすることが困難である。そこで EM 法を用いることで効果的な推定が可能となる。EM 法は、各要員が確率密度関数で、現象は複数要因が混合した、混合モデルを想定している。ただし線形和および混合比を固定すると仮定する。EM 法で推定する対象は2つあり、各要因のモデルのパラメータと混合比である。EM 法の計算原理は各EMstep を最尤推定で算出することにある。ここでの推定値は推定値の極限である。Expectation Step と、Maximization Step から構成されており、Estep (期待値計算)では推定するパラメータをθ とした時、期待値を求める step である。次式で表すことができる。

$$Q(\theta;\theta') = \sum_{\boldsymbol{x}^{(i)}\in D}\sum_{c} P(c|\boldsymbol{x}^{(i)};\theta')\log P(c,\boldsymbol{x}^{(i)};\theta)$$

Mstep ではθの最尤推定を行う。次式で表すことができる。　　$\theta^{max} = \arg\max_{\theta} Q(\theta;\theta')$

Estep で事後確率の期待値を計算する際、多次元正規分布を使うことで、多くの変量の関係がモデル化でき、混合モデルや複雑な現象でも記述できる。これは、変数間の関連を分散共分散行列として表現することで効果的な推定が可能になるからである。しかしここで膨大な計算を必要とする。例えば、100 人の生徒のテストの成績を EM で解析する際には、EM タスクの共分散の再計算を1回行うだけで、生徒一人一人が取りうる値全てに対し確率密度関数を演算するため、最低でも1万回確率密度関数を計算する必要がある。このことが計算効率を低下させ、処理精度の向上が課題となる。

3. モンテカルロ法

モンテカルロ法の基本は確率密度関数p(X)に従う乱数の平均{$\sum f(X)/N$}は期待値$E_{p(x)}[f(x)]$に近似できることであり、確率分布$\int P(X)$に従う乱数を生成できることである。次式で表せる。

$$E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f\left(x^{(i)}\right)$$

当該の生徒集合の成績に対応する統計量も的確に生成できる。

4. 提案手法

本研究では、処理性能の向上を目的とし EM 法にモンテカルロ法を利用する。計算過程で、Estep で再計算した確率値を使ってMstep(最尤推定)で平均と分散共分散を計算する。訓練データの確率をMstepごとに繰り返し計算することが計算処理を大幅に低下させる理由である。Estep を与えられた事象のモデルからサンプリングを行うタスクに置き換え、Mstep では,Estep でサンプリングした値を訓練データに追加することで、パラメータを再計算し、モデルの再構築を行うタスクに置き換える。実験手順を、例を用いて表す。訓練データを 1000 人の学校の生徒の英語、数学の成績がわかるとする。ただ生徒は、利き手が左か右に別れており、700 人はどちらが利き手か判明し、残り 300 人は不明とする。700 人の訓練データを各利き手ごとに分け、サンプリングして得た値を追加した701人データに拡張、もう一度必要とするパラメータを計算し701人のデータから702人のデータに拡張する。これを繰り返し、700人を3000人にデータを増やしていく。この時、元とする訓練データを読み込むのは最初の一度限りである。3000 の成績データを利用

し所属尤度を計算し、残り 300 人の利き手のクラス推定をする。サンプリングの具体的な手順を示す。Step1:訓練データ D より平均ベクトルμ及び,分散共分散行列∑生成する。Step2:$\sum = AA^t$を満たすAを固有値分解で抽出。Aは固有値Λ,固有ベクトル P とすると次式で求められる。

$$AA^t = P\sqrt{\Lambda}\sqrt{\Lambda}P^t$$

Step3 ： 標準正規分布からサンプリングしたN次元の乱数$Z\sim N(0,1)$を生成。Step4 ： サンプリング値 X を求める。X は次式で求めることができる。

$$X = P\sqrt{\Lambda}Z+\mu = AZ + \mu$$

Step5 ： 訓練データ D に X を追加し、訓練データ D'として再構築を行う。Step6 ：訓練データ D'よりμ及び,∑を再生成し、Step2 に戻り繰り返しサンプリングを行う。ここで共分散行列を分解し、変数間の重みとして表現することで、訓練データの変数間の関係を反映した値をサンプリングする。

5. 実験

5.1 実験準備

本研究では，提案する手法の処理精度での優位性を示すため、従来の EM 法と精度および同等の精度を出力するまでの計算時間について比較を行う。精度は、クラス分類の正答率について比較を行うものとする。従来の EM 法の収束条件は、対数尤度計算を行い、偏移が 0.001 以下となった時点で収束とし実験を行う。収束後のクラス分類の精度を提案手法の精度との比較に利用するものとする。UCI Machine Learning Repository よりワインの成分データ 178 件を使用。変数は[Alcohol],[Alkalinity of ash],[Flavonoids],[color intensity]の４変数、３種の葡萄品種の 3 クラスで訓練データ D を構成する。この時、次元間の相関係数は Alkalinity of ash と Color intensity の間 0.001 を除いて、0.1 以上 0.5 未満である。訓練データと比較用データに 7:3 の比率でランダムに分け、同作業を繰り返し 20 のデータセットを作成、それぞれに対しクラス正答率、計算時間を求め,平均を取り最終的な値とする。ここで提案手法では、比率 7 の訓練データにはあらかじめ葡萄の品種のラベルを付与し、それぞれさんクラスに分割する。分割したそれぞれのデータに対し、同時に学習させる。共分散および固有値分解を計算するために、統計解析ツール R の行列計算関数を用いて行う。実験に必要なアルゴリズムも同様に R 言語で作成する。

5.2 実験結果

従来の EM 法と提案手法のサンプリング数ごとのクラス分類の精度および計算時間を以下の表に示す。また従来の EM 法でのクラス分類の精度は最尤原理で事象の所属クラスをラベル付し、提案手法では事象の相対確率を尤度計算に用いて行う。

サンプリング数	正答率(%)	実行時間(s)
従来のEM	84.1666667	8.4272
0	63.9814815	0
300	73.1481481	0.99085
400	79.2592593	1.3661
500	81.6666667	1.66445
600	81.7592593	1.96235
700	85.1851852	2.25635
800	83.9814815	2.69745
900	85.462963	3.1163

表１　クラス分類精度と実行時間

従来の EM 法での分類精度は 84%、提案手法がこの精度を超える時点は 700 回のサンプリング数の時点である。この時、従来の EM の計算時間は 8.42(s)に対し、提案手法は 2.25(s)である。

サンプリング数	正答率		
0	64%	2000	87%
300	73%	3000	89%
400	79%	4000	87%
500	82%	5000	88%
600	82%	6000	89%
700	85%	7000	90%
800	84%	8000	89%
900	85%	9000	89%
1000	85%	10000	89%

表 2 サンプリング数と精度

サンプリング数が 3000 回を超えた時点から精度の変動が横這いとなる。

5.3 考察

サンプリング回数と精度の偏移について、ある時点で正答率の変動が収まる。原因としては、訓練データ D'がサンプリングを重ねるごとに次第に密になり、疎である範囲もしくは、D'を大きく変化させるような特出した値が出なくなるため訓練データ D'が大きく変化しないものと考える。実際、変数 1 のアルコールを例にサンプリング回数ごとの分散をまとめると、3000 回未満での最大の分散の変化は 0.0099 で 3000 回以上での変化はおよそ 1/3 の 0.0034 である。よってありきたりのようなデータしかサンプリングしなくなるのではないかと考えられる。

6. 結論

従来の EM 法の処理精度は 84.16%のクラス正答率で計算時間は 8.42 秒,今回の提案手法は 700 回のサンプリングで従来の EM 法と同等の精度 85.1%となり、計算時間は 2.25 秒となった。計算時間が 2/7 で約 73%の減少となった。今後の課題として、今回、1 桁の次元での実験を行ったが、次元数とデータ数を大幅に拡張した状況下で処理性能を優位に行えるか試行する必要がある。

参考文献

[1] 離散型確率分布を用いた EM 法，電子情報通信学会総合大会 学生ポスターセッション，広島大学，東広島 ,共著(五十嵐 有真，三浦 孝夫)，平成 32 年 (2020) 3 月