ISS-A-071

離散型確率分布を用いたEMアルゴリズム

五十嵐　有真　　三浦　孝夫
法政大学理工学部創生科学科

1. 前書き

本稿では、確率分布関数を想定しない表形式確率分布推定のためのEMアルゴリズムを提案する。基本的なEMアルゴリズムは、(以下、関数型 EM アルゴリズム)離散値や確率分布に従うか未知の値に関しては、関数を与えることができないため、従来の計算方法では機能しない。提案手法である表形式型 EM アルゴリズムでは、確率分布関数を定義せず、表を使って E,M ステップを計算する。データがどのような確率分布に従っているかという知識は一般に不明である。また判明する場合であっても、これを仮定しないことに利点がある。

2. 提案手法

表形式型EMアルゴリズムでは確率分布関数を定義せず、表を使ってEステップMステップを計算する。Eステップでは、階段関数によって近似した分布自体を定義する。学習データ X(表１)を出現頻度でベクトル化する(表 2)。出現確率 P(X)に変換し(表 3)、累積分布の表を作成し(表４)階段関数自体を定義する。この一連の流れをEステップとする。Mステップでは、サンプリングによる表の細分化と分布の再構成を行う。一様乱数=0.4 が発生した場合、表４のi番目に 0.4 を追加する(表5)。0.4 を挿入した気温データの前後の尤度比を e(i-1)e(i+1)に対応させ、表を分割し、e を算出する。これを学習データ X もしくは、前ステップの開発データ En-1 に追加して、開発データ E n を作成する(表 6)。

表 1 学習データ X

学習データ(初期値)	15	16	18	…	20	21	22

表 2 データ X のベクトル化

	15	16	18	…	20	21	22
気温データ	1	2	3		2	1	1

表 3 出現確率 P(X)

	15	16	18	…	20	21	22
気温データ	0.1	0.2	0.3	0	0.2	0.1	0.1

表４ 累積確率分布表 ΣP(X)

	15	16	18	…	20	21	22
気温データ	0.1	0.3	0.6		0.8	0.9	1

表 5 逆関数計算表

	15	16	e	18	…	20	21	22
気温データ	0.1	0.3	0.4	0.6		0.8	0.9	1

表 6 開発データ En

開発データ(1回目)	15	16	16.666	18	…	20	21	22

再び E ステップに戻り、階段関数分布の再構築を行う。このように確率分布関数を定義せず、離散値の表のみでEMアルゴリズムを実行する。

3. 実験

3-1. 実験手順

表形式 EM アルゴリズムを 1000 回から 10000 回まで 1000 刻みで試行する。クラスの正答率をそれぞれの試行回数で確認し評価する。尤度比の推移を追跡し、収束しているか判定を行う。データは気象庁の気象データから6月１日～6月30日の1日の平均気温一ヶ月分の気温のデータを使用する。学習データ数は 270 件である。

表 7 学習データ初期値

新潟クラス	鹿児島クラス	北海道クラス
16.2	21	9.4
16.4	21.4	9.7
16.5	21.4	9.9
16.6	21.4	10
16.6	23.6	10
16.8	23.6	10.1
17.4	23.9	10.4
17.4	23.9	10.5
17.4	24.1	11
17.6	24.1	11
17.6	24.2	11.2

表 8 開発データ

新潟クラス	鹿児島クラス	北海道クラス
16.2	21	9.4
16.20389	21.28403	9.40057
16.20428	21.28972	9.40087
16.20456	21.2916	9.40218
16.20512	21.29174	9.40308
16.2054	21.29181	9.41463
16.20589	21.29188	9.42586
16.2134	21.29239	9.43001
16.21343	21.2994	9.43042
16.21345	21.30786	9.43873
16.21619	21.33296	9.44842
16.21658	21.32366	9.44898
…	…	…

3-2. 実験結果

表形式型 EM アルゴリズムの結果を表 8 に示す。
3 クラス分類の正答率をそれぞれ計算する。1000 回から 10000 回まで線刻みそれぞれの正答率を表 9 に記す。

表 9 クラス分類正答率

試行回数	1000回	2000回	3000回	4000回	5000回
正答率(%)	80.7692308	80.7692308	78.2051282	80.7692308	82.0512821
	6000回	7000回	8000回	9000回	10000回
	80.7692308	84.6153846	80.7692308	80.7692308	82.0512821

最も正解率が高かったのは 7000 回で 84.6%の正答率である。収束条件を満たした回を 1000 回刻みで表 10 に記す。

表 10 収束状況

試行回数	1-5000	6000-6999	7000-7999	8000-8999	10000-10999
収束条件を満たした回数	0	5	18	3	21

初めて収束状況が確認できたのは、6409 回である。

4.考察

最も 7000 回の時点での正答率が最も高い。また最も低いのは 3000 回のである。この最も正答率が高い回数と低い回数で、正解と誤答の尤度の差を比較する。

表 11 誤答と正答との尤度の差

	3000回	7000回
差が最小の誤答	0.000000841	0.00014132

表 11 から 3000 回の時は微細な差で誤答したことがわかる。回数が増えるごとにこの差が埋まり正答率が向上したと考えられる。初期値の依存が大きいため、上限下限を超えるデータに対して正答率が著しく低下する。

5.結論

本実験において、EM として機能し、クラス分類の精度が84%の分類器を得た。6409 回を超えると、収束条件を満たすことが確認できた。表形式確率分布推定のための EM アルゴリズムが立証できた。

参考文献

[1] Jeff Bilmes , A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, ICSI Technical Report ,1998 1-13