能動学習に基づく自律的 EM 学習

井上　眞乙†　　三浦　孝夫†
†法政大学理工学部創生科学科

1. EM アルゴリズムと能動学習

本研究では，能動学習に基づいた EM アルゴリズムを用いる新聞記事の文書分類を行う[1].

分類器は，データを予め決まったカテゴリに仕分ける機構であり，構築には学習データを用いる。少量の学習データで高精度の分類器を構築するため，EM アルゴリズムを用いて分類する．更に高性能分類器を得るため，能動学習を EM アルゴリズムの E ステップと M ステップの間に組み込む．EM アルゴリズムが自律的に外部から改めて修正・学習を繰り返すことで，収束速度の改善と高い分類精度の実現をする．データ選定のため，最大・最小確信度法を用いる．閾値を設け，選定されたデータの語頻度が一定の影響力を持つことを保証する．収束速度改善のための計算では，対数尤度期待値の計算を行う．

2. 実験と考察

EM アルゴリズムと最小確信度および最大確信度で選定する能動学習を用いる EM アルゴリズムの 3 手法で所属確率を推定し，適合率および収束速度を比較する．実験では，「毎日新聞 2012 データ集」1 月 1 日の先頭からスポーツ記事 50 件，経済記事 50 件，芸能記事 50 件を対象とする．適合率には MicroPrecision を用いる．学習データとしてスポーツ記事，経済記事，芸能記事に 90 件ずつ割り当て，テストデータとしてスポーツ記事，経済記事，芸能記事から各 50 件を用いる．

EM では収束回数が 6 回，全体の適合率は 121/150=80.7%，MicroPrecision では 87.3%である．Largest-ALEM では能動学習によりテストデータから選定された記事データは，スポーツ記事から 2 件芸能記事から 2 件である．収束回数は 8 回，全体の適合率は 90/146=61.64%, MicroPrecision では 90.4%である. Least-ALEM では能動学習により選定された記事データは，スポーツ記事から 1 件，経済記事から 1 件，芸能記事から 1 件の合計 3 件である. 収束回数は 7 回, 全体の適合率は 103/147=70.1%, MicroPrecision では 89.8%である．LargesrALEM と比べ、経済クラス記事の正しい割り当てが大幅に改善している．これは分類境界上のデータを強制的に学習することで、精度を改善させることができることを示している．

全体の適合率: 87.3(%)　収束回数: 6(回)

結果 / 正解	スポーツ (件)	経済 (件)	芸能 (件)	記事数 (件)
スポーツ	31	4	15	50
適合率(%)	62			
経済	0	42	8	50
適合率(%)		84		
芸能	0	2	48	50
適合率(%)			96	
記事数合計(件)	31	48	71	150

選定データ: スポーツ記事2件(記事番号: 327,330)
芸能記事2件(記事番号: 5928,5608)
全体の適合率: 90.4(%)　収束回数: 8(回)

結果 / 正解	スポーツ (件)	経済 (件)	芸能 (件)	記事数 (件)
スポーツ	43	0	5	48
適合率(%)	89.6			
経済	8	0	42	50
適合率(%)		0		
芸能	1	0	47	48
適合率(%)			97.9	
記事数合計(件)	52	0	94	146

選定データ: スポーツ記事2件(記事番号: 327,330)
芸能記事2件(記事番号: 5928,5608)
全体の適合率: 90.4(%)　収束回数: 8(回)

結果 / 正解	スポーツ (件)	経済 (件)	芸能 (件)	記事数 (件)
スポーツ	43	0	5	48
適合率(%)	89.6			
経済	8	0	42	50
適合率(%)		0		
芸能	1	0	47	48
適合率(%)			97.9	
記事数合計(件)	52	0	94	146

選定データ: スポーツ記事1件(記事番号: 355)
芸能記事1件(記事番号: 2022)
経済記事1件(記事番号: 807)
全体の適合率: 89.8(%)　収束回数: 7(回)

結果 / 正解	スポーツ (件)	経済 (件)	芸能 (件)	記事数 (件)
スポーツ	34	0	15	49
適合率(%)	69.4			
経済	0	32	17	49
適合率(%)		65.3		
芸能	0	2	47	49
適合率(%)			95.9	
記事数合計(件)	34	34	79	147

3. 結論

本研究では，能動学習による有用なデータの選定方法を提案した．この能動学習を，EM アルゴリズムの E ステップと M ステップの間に組み込むことで，自律的にデータ選定する ALEM を提案した

参考文献

[1]A. K. McCallum, K. Nigam, “Employing EM and Pool-Based Active Learning for Text Classification”, ICML98, pp. 350-358