ISS-P-59

最大エントロピーモデルのための素性選択の簡素化

後藤 仁† 三浦 孝夫†
† 法政大学理工学部創生科学科

1. 最大エントロピー法と分類器

本研究では新聞記事コーパスを用いて最大エントロピー法でモデルを推定し分類器を構築する．構築にあたって語の共起から同時関係を考慮した分類を行うための素性を選択する．本研究では,語の頻度を用いて同時関係の前提部にあたる語を抽出し素性とする.

2. 素性選択とその簡素化

本研究では，語Xが同時関係X ⇒ Yの前提部Xであるかを判定する．Nは文書件数，sは最少支持度，cは最少確信度，supp(X)は語Xの頻度，supp(X ∪ Y)は集合{X, Y}の頻度を表す．まず同時関係X ⇒ Y (s, c)が成り立つための条件を以下に示す.

$$c < \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \quad (1)$$

$$sN < supp(X) \leqq \mathrm{supp}(X \cup Y) \quad (2)$$

(1)は同時関係の信頼性，成り立つ割合を示す.極大頻出項目集合の頻度は自身の部分集合に比べて最もsNに近い．極大頻出項目集合の頻度supp(X ∪ Y)が最少出現回数sNに最も近いことから，supp(X ∪ Y)をsNで置き換える．X ∪ Yの生成を不要にし前提部Xを抽出する点がアイデアである．要約すると以下のようになる

(1) 極大頻出集合の頻度が最少頻度に最も近い
(2) 極大頻出集合の頻度を最少頻度とみなす
(3) 最少頻度<語 X の頻度 <最少頻度/最少確信度
(4) (3)の条件を満たせば語Xは同時関係の前提部と判断し素性とする

3. 実験

実験にはCD-毎日新聞2013年度11月の記事を用いる．7296件のうち，11月1日から11月20日までの記事4550件を学習データとし，11月21日から11月30日までの記事2746件をテストデータとする.素性選択では最少支持度 s=0.1，最少確信度 c=0.8 の下で式(3)の条件を満たす語をクラス毎に抽出し，素性に用いる語とする．ベースラインは素性を極大頻出項目集合とする ACME で，最少支持度 s=0.1 でクラス毎に抽出した極大頻出項目集合を素性に用いる．極大頻出項目集合の抽出にはLCM v5.3[2]を用いた.評価の指標としてクラス毎の再現率(Recall)，精度(Precision)と両者の調和平均をとったF値と素性選択に要した実行時間を用いる.

4. 実験結果

		再現率(Recall)		精度(Precision)		F値	
	記事数	提案手法	ﾍﾞｰｽﾗｲﾝ	提案手法	ﾍﾞｰｽﾗｲﾝ	提案手法	ﾍﾞｰｽﾗｲﾝ
1面	172	0.064	0.413	0.139	0.243	0.088	0.306
2面	111	0.054	0.018	0.125	0.095	0.075	0.030
3面	44	0.136	0.364	0.545	0.500	0.218	0.421
ｽﾎﾟｰﾂ	581	0.967	0.824	0.857	0.735	0.909	0.777
国際	223	0.578	0.386	0.629	0.551	0.603	0.454
家庭	123	0.285	0.260	0.486	0.525	0.359	0.348
文化	16	0.063	0.188	1.000	0.750	0.118	0.300
特集	49	0.000	0.020	-	0.500	-	0.039
社会	569	0.724	0.703	0.504	0.419	0.594	0.525
社説	185	0.465	0.454	0.389	0.506	0.424	0.479
経済	291	0.667	0.515	0.595	0.620	0.629	0.563
総合	279	0.430	0.179	0.430	0.373	0.430	0.242
芸能	61	0.311	0.295	0.760	0.818	0.442	0.434
解説	16	0.000	0.000	0.000	-	-	-
読書	26	0.077	0.154	0.500	0.571	0.133	0.242

分類結果を示す．提案手法は15クラス中8クラスでF値がベースラインより高い．再現率でも15クラス中9クラスでベースラインを上回るが，精度は15クラス中6クラスで上回る．提案手法では特集・解説クラスの正解が0件で，ベースラインでも解説クラスの正解が0件となる．提案手法・ベースライン共に最も再現率とF値が高いのはスポーツクラスで，提案手法では再現率0.967，ベースラインでは0.824となる．社説クラスでは提案手法がベースラインを再現率で上回るものの，精度とF値で下回っている．1面クラスの再現率が提案手法では0.064，ベースラインは0.413と差がある.

5. 結論

本研究では最大エントロピーモデルのための素性選択のため同時関係を用いる手法を提案した．集合素性と比較して，分類のF値は15クラス中8クラスでベースラインよりよい結果を与えた.

参考文献

[1]Risi Thonangi, Vikram Pudi ACME:An Associative Classifier based on Maximum Entropy Principle, 2005
[2]Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura LCM:An Efficient Algorithm for Enumerating Frequent Closed Item Sets 2003
[3]濱崎邦秀，三浦孝夫，“最大エントロピーモデルとデータマイニングを用いた多重ラベル分類”, DEIM Forum 2015