新聞の社説を教師信号とする文章の右翼度・左翼度判定 第２報

畑中充宏 [1] **金丸敏幸** [2] **村田真樹** [3] **掛谷英紀** [4]

筑波大学 [1]

情報通信研究機構 [2]

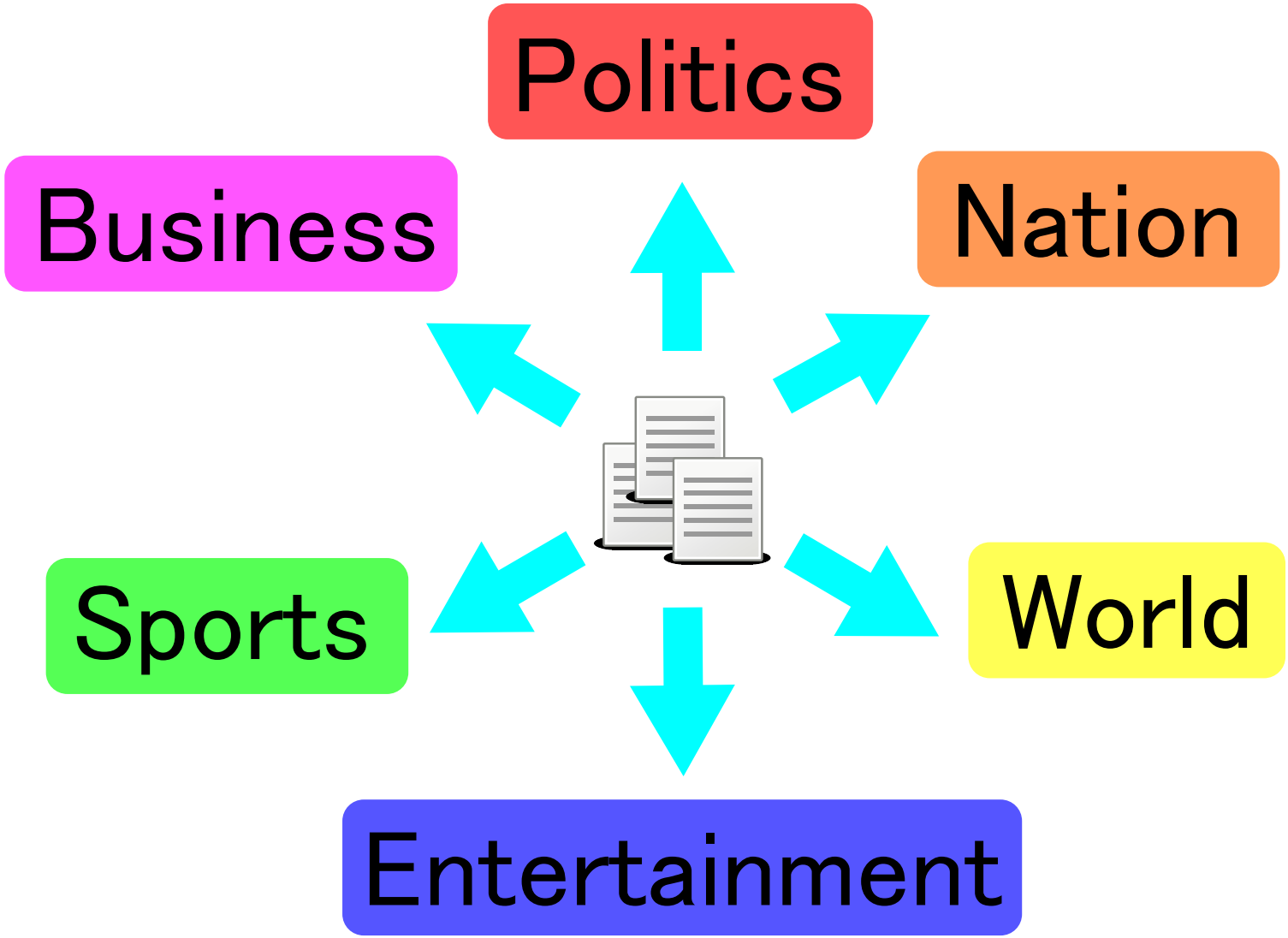
情報通信研究機構 [3]

筑波大学 [4]

September 15, 2008

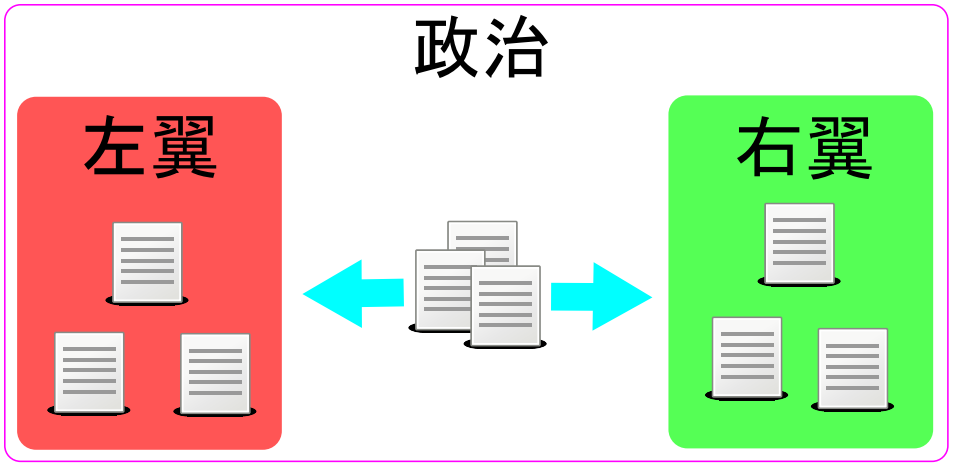
Outline

従来の文書分類



政治的イデオロギーによる文書分類

- **政治的イデオロギーを分類する研究はほとんど行われていない**
  - Why? 政治的イデオロギーの指標が得にくい
- 社説を政治的イデオロギーの指標にする
- イデオロギーにしたがって文書分類できる

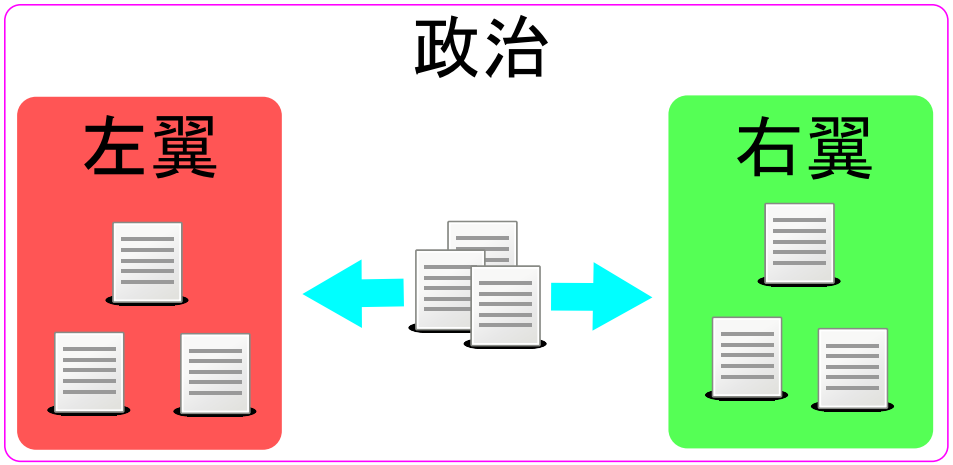


政治的イデオロギーによる文書分類

- **政治的イデオロギーを分類する研究はほとんど行われていない**

  Why? **政治的イデオロギーの指標が得にくい**

- **社説を政治的イデオロギーの指標にする**
- **イデオロギーにしたがって文書分類できる**

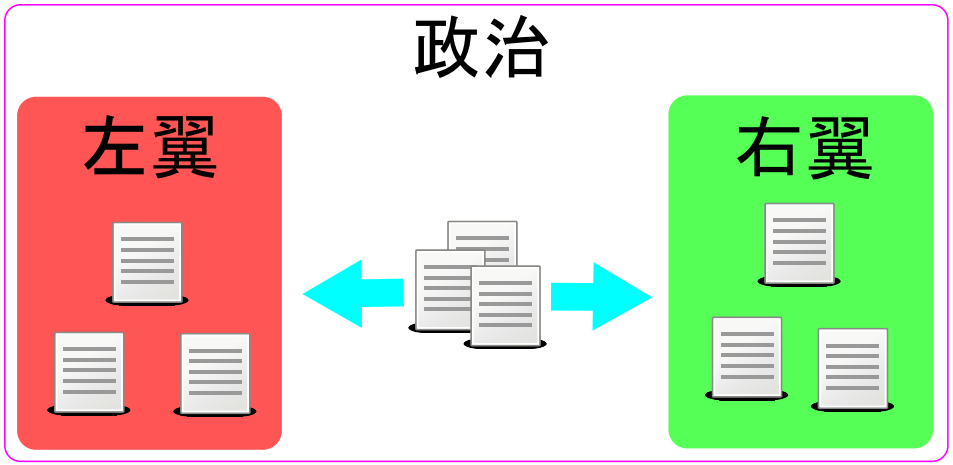


政治的イデオロギーによる文書分類

- 政治的イデオロギーを分類する研究はほとんど行われていない

  Why? 政治的イデオロギーの指標が得にくい

- 社説を政治的イデオロギーの指標にする
- イデオロギーにしたがって文書分類できる

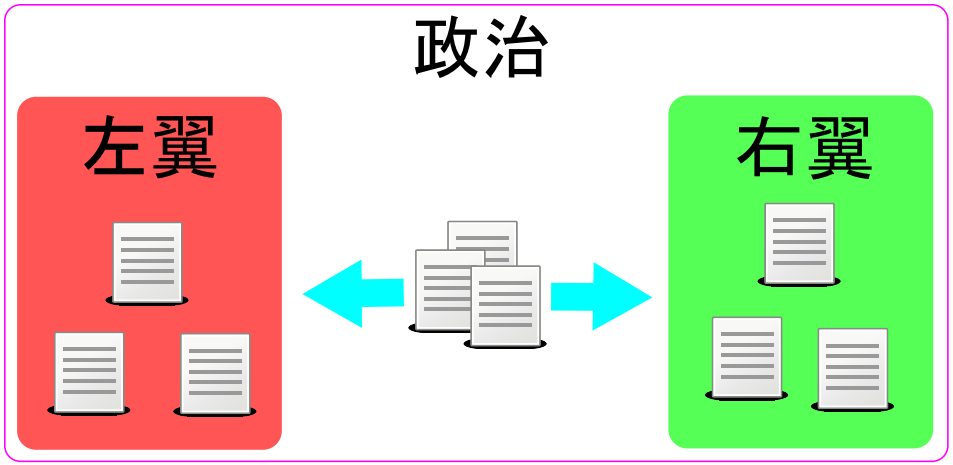


政治的イデオロギーによる文書分類

- 政治的イデオロギーを分類する研究はほとんど行われていない

  Why? 政治的イデオロギーの指標が得にくい

- 社説を政治的イデオロギーの指標にする
- イデオロギーにしたがって文書分類できる



判定の準備

教師用信号

右翼系 読売新聞

左翼系 毎日新聞

用いるデータ

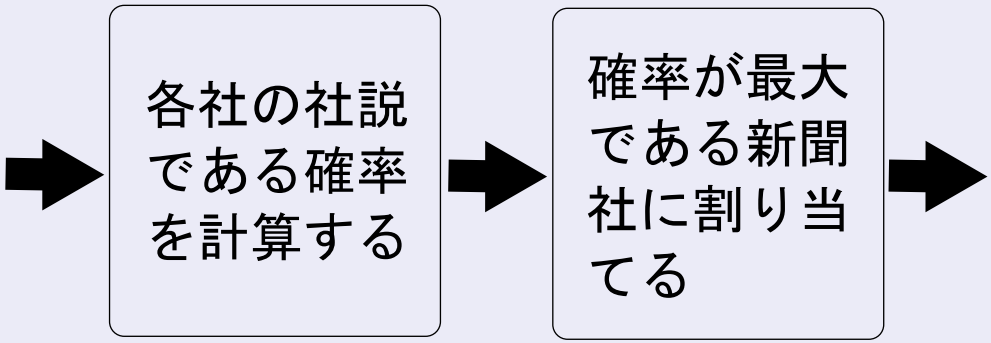
単語	名詞・動詞
熟語	名詞が 2 つ以上連なったもの・形容詞の係る名詞
末尾表現	句点から数えて 3〜7 文字以内

学習方法

最大エントロピー法

社説の判定・行う実験

判別の流れ

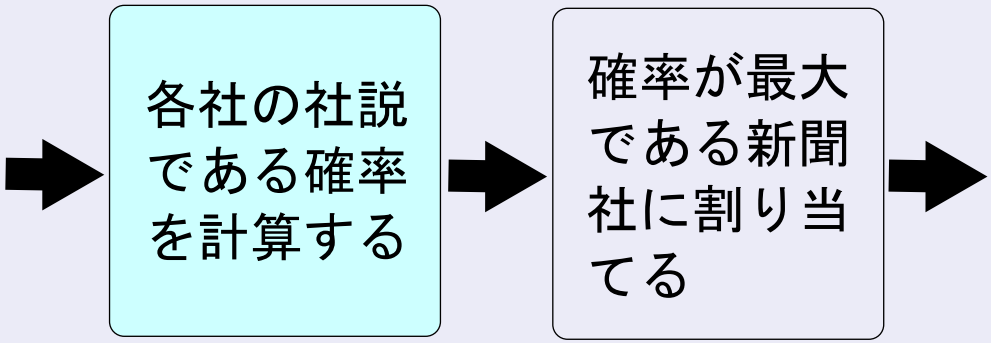


行う実験

1. 毎日新聞と読売新聞の社説を用いてクロスバリデーションで実験
2. 朝日新聞と産経新聞と日本経済新聞の社説をテストデータとして実験
3. 学習データを読売・毎日・日経の 3 社にして実験
   1. 10 分割のクロスバリデーションで実験
   2. 朝日新聞・産経新聞をテストデータとして実験

社説の判定・行う実験

判別の流れ

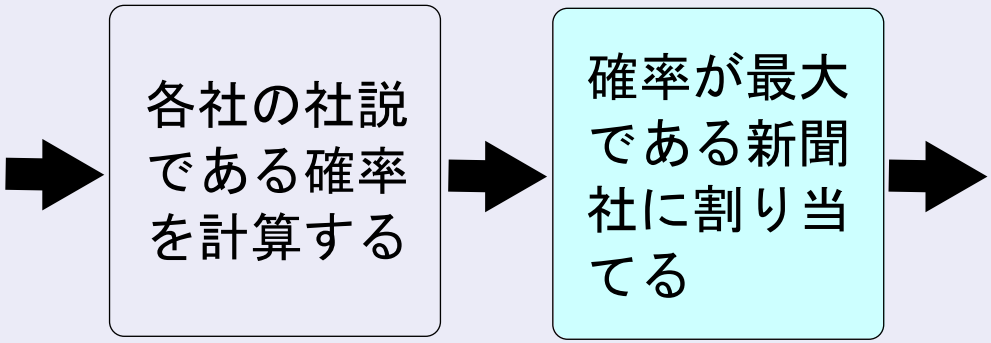


行う実験

1. 毎日新聞と読売新聞の社説を用いてクロスバリデーションで実験
2. 朝日新聞と産経新聞と日本経済新聞の社説をテストデータとして実験
3. 学習データを読売・毎日・日経の 3 社にして実験
   1. 10 分割のクロスバリデーションで実験
   2. 朝日新聞・産経新聞をテストデータとして実験

社説の判定・行う実験

判別の流れ



行う実験

1. 毎日新聞と読売新聞の社説を用いてクロスバリデーションで実験
2. 朝日新聞と産経新聞と日本経済新聞の社説をテストデータとして実験
3. 学習データを読売・毎日・日経の 3 社にして実験
   1. 10 分割のクロスバリデーションで実験
   2. 朝日新聞・産経新聞をテストデータとして実験

実験 1(毎日・読売)

学習データ

15 年分の毎日新聞と読売新聞の社説

実験方法

10 分割のクロスバリデーション

結果

- **正解率** 91.7%
- 高い確信度で正解している

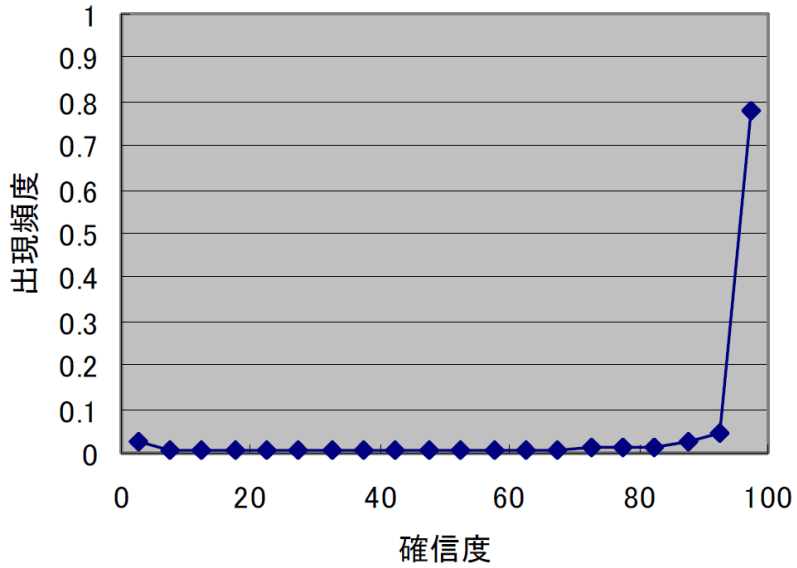


Figure: 社説の判定結果の分布

実験 1(毎日・読売)

学習データ

15 年分の毎日新聞と読売新聞の社説

実験方法

10 分割のクロスバリデーション

結果

- 正解率 91.7%
- 高い確信度で正解している

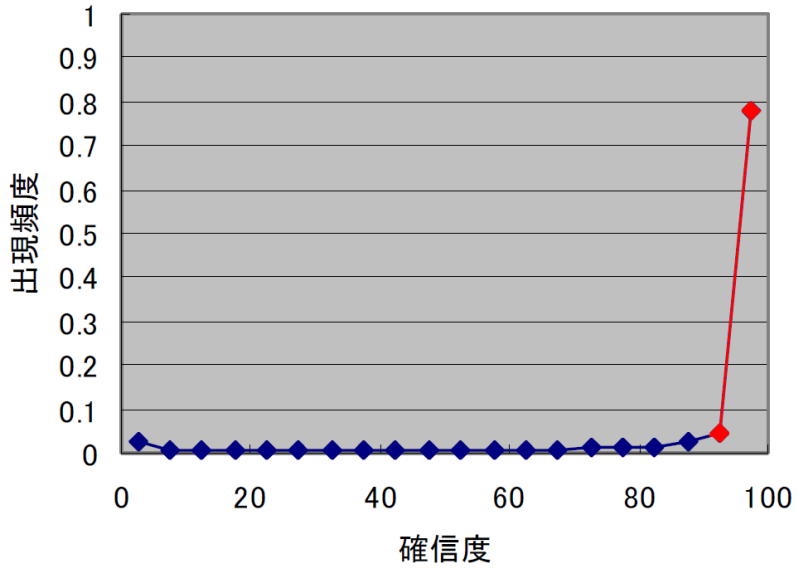


Figure: 社説の判定結果の分布

実験 2(朝日・産経・日経)

テストデータ

- 1 年分の朝日新聞の社説
- 4ヶ月分の産経新聞の社説
- 1 年分の日本経済新聞の社説

変更点

1. 素性データを単語・熟語のみにする
2. 数字を含む素性を削除

実験 2 の結果 I

朝日新聞

Table: 朝日が「毎日」と判定された割合

	条件なし	条件 1	条件 2
朝日 2006	86.7%	86.9%	68.1%
朝日 2007	92.3%	90.9%	74.9%

- どの条件でも左翼系の新聞社である毎日新聞と判定
- 右翼・左翼度判定システムとしては望ましい

実験 2 の結果 II

産経新聞

Table: 産経が「読売」と判定された割合

	条件なし	条件 1	条件 2
産経 2007	25.0%	36.5%	62.5%

- 条件なしでは 75.0%の確率で毎日新聞と判定
- 条件 1 でも改善せず
- 条件 2 で 62.5%の確率で右翼系の新聞社である読売新聞と判定

実験 2 の結果 III

日経新聞

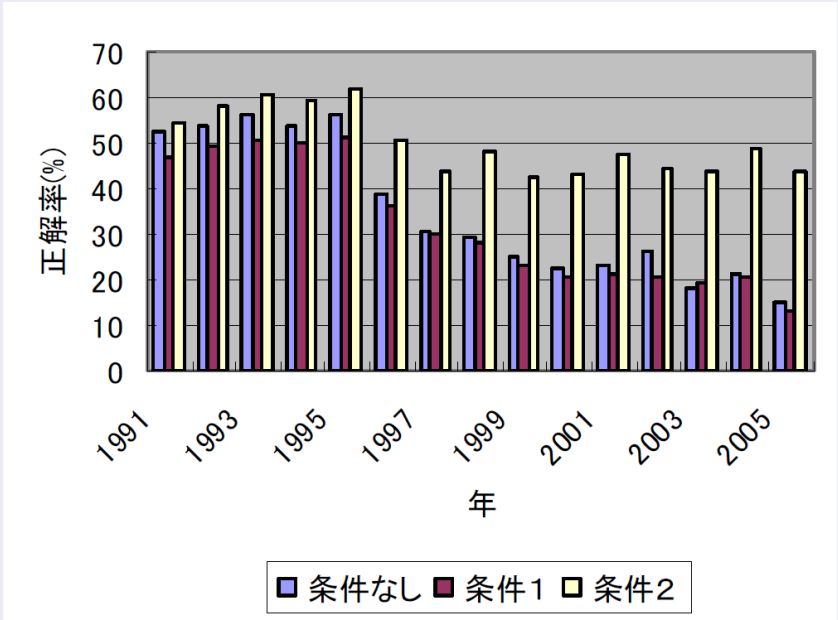


Figure: 日経が「毎日」と判定された割合

- 1995 年までは 50%付近をさまよう
- 1996 年以降は読売新聞に近いと判定される
- 条件 1 では改善は見られない
- 条件 2 では 50%付近にとどまっている

実験 3

学習データ

- 3 社についての 15 年分の社説
- 数字を含む素性の排除

テストデータ

- 1 年分の朝日新聞の社説
- 4ヶ月分の産経新聞の社説

Table: 朝日・産経の判定結果

	読売	毎日	日経
朝日 2006	26.1%	55.7%	18.1%
朝日 2007	22.0%	60.6%	17.4%
産経 2007	31.5%	26.0%	43.0%

- クロスバリデーションの正解率は 83.3%
- 朝日新聞で最も高く，産経新聞で最も低い毎日新聞は良い教師信号
- 読売新聞・日経新聞は互いに素性を食い合っている

考察

- 最大エントロピー法は，どの素性がデータを判定するのに重要になるかを示す変数 α が算出される
- α 値の高い素性は，思想を反映するものと新聞社の表記の違いが影響しているものがある
- 数字は表記が複数あり，必ず存在するため影響が大きい

思想を反映

- 国際社会
- 市場経済化
- 庶民
- キム

表記の違い

- こたえる
- 応える
- 小泉首相
- 小泉純一郎首相
- 3
- 三

考察

- 最大エントロピー法は，どの素性がデータを判定するのに重要になるかを示す変数 α が算出される
- α 値の高い素性は，思想を反映するものと新聞社の表記の違いが影響しているものがある
- 数字は表記が複数あり，必ず存在するため影響が大きい

思想を反映

- 国際社会
- 市場経済化
- 庶民
- キム

表記の違い

- こたえる
- 応える
- 小泉首相
- 小泉純一郎首相
- 3
- 三

まとめ

- 新聞の社説を教師信号として文書を判定システムを提案
- 読売新聞・毎日新聞では高い正解率で判定可能
- 他の新聞社に関しても，右翼・左翼の判定可能
- 思想を反映しない素性を排除すると，判定結果は向上
- 学習データを増やせば，さらに正確な判定が期待できる

最大エントロピー法

事象 t と h が同時に出現する頻度 $O(t,h)$ から条件付き確率 $P(t|h)$ を推定するアルゴリズム
式 1 の制約を満たしつつ，式 2 を最大化するようなパラメータを推定する

$$\forall f_i \sum_{t,h} \hat{P}(h) P(t|h) f_i(t,h) = \sum_{t,h} \hat{P}(t,h) f_i(t,h) \tag{1}$$

$$E(P) = -\sum_{h} \hat{P}(h) \sum_{t} P(t|h) \log P(t|h) \tag{2}$$

クロスバリデーション (交差検定法)

- N 割して，$N-1$ 個のグループを訓練に使い，1 個を評価に使う．
- データが限られている場合に使われる

Figure: 10 分割のクロスバリデーションのイメージ