3D1-2

人間の因果的直感を用いたスパム分類器

Spam classifier using causal intuition of human

清水 隆宏[*1] 大用 庫智[*1] 高橋 達二[*2]
Takahiro Shimizu Kuratomo Oyo Tatsuji Takahashi

[*1] 東京電機大学大学院 Graduate School of Tokyo Denki University
[*2] 東京電機大学理工学部 School of Science and Technology, Tokyo Denki University

A model that implements cognitive biases such as symmetry and mutual exclusivity has been proposed. The model, LS (loosely symmetric) formula, faithfully describes the causal intuition of humans. We have shown that LS, operated as a value function in the framework of reinforcement learning, shows versatile efficiency. In this study, we utilize LS in spam mail filtering. The efficacy of intuitive judgment of causal relationship of human that has been forged through the course of evolution is tested.

1. はじめに

人間の因果的直感を忠実にモデリングし，対称性バイアスや相互排他性バイアスといった認知バイアスを実装したモデル（Loosely Symmetric model: LS）が提唱され，機械学習分野での諸タスクにおいて優れた成績を収めている．

本研究ではスパムメール分類において，LS を用いることにより，進化の過程で鍛えられた人間の因果関係に関する直観的判断を持つ分類器が有用であるかを検討する．

2. Loosely symmetric model

緩い対称性モデル (LS) とは，人間の因果帰納等に存在する“ 対称性バイアス ”および“ 相互対称性バイアス ”という 2 つの非論理的な認知バイアスを緩やかに持つ確信度のモデルである[篠原 2007].

原因となる事象 p と結果となる事象 q があるとき，対称性バイアスは「$p \rightarrow q$」という情報からから「$q \rightarrow p$」を導き，相互排他性バイアスは「$p \rightarrow q$」から「$\bar{p} \rightarrow \bar{q}$」を導く傾向を表す．これらは論理学において逆と裏の関係に当たり論理的には誤りである．しかし，人間は因果帰納において度々このような推論を行う事が知られている．人間が常にこのようなバイアスを働かせているとは考え難い．LS は地の不変性などを用いてこれらバイアスを柔軟に変化させる事により人間の因果帰納実験に対して高い相関を持つ事が示されている．

本章で述べるモデルにおける a, b, c, d はそれぞれ p と q という事象の共起頻度，あるいは共起確率である pq，$p\bar{q}$，$\bar{p}q$，$\bar{p}\bar{q}$ に対応する(表 1).

表 1 ： 共起情報の 2×2 分割表

	q	$\bar{q}$
p	a	b
$\bar{p}$	c	d

$$LS(q|p) = \frac{a + \left(\frac{b}{b+d}\right)d}{a + b + \left(\frac{a}{a+c}\right)c + \left(\frac{b}{b+d}\right)d} \quad (1)$$

3. 提案手法

提案手法では，トレーニングデータから Word と SPAM, NOT SPAM の共起頻度として保持する．

スパムメールの判別を行うために，教示として与えられるメールデータを LS が扱える形に変換する．表 2 は判別を行うメール本文に存在する Word を共起頻度から抽出し，Word と SPAM, HAM の共起情報を表 2 のように分割する．a, b, c, d の計算は式(2)(3)(4)(5) の通りである．求めた a, b, c, d を用いて確信度の計算を行う．

得られた各 Word の SPAM, NOT SPAM の確信度の大小関係を比較しカウントを行う．全ての Word で比較処理を行いカウントした数が多い選択肢により SPAM, NOT SPAM か判別を行う．

表 2 ： 抽出した共起頻度表

	$SPAM$	$NOT\ SPAM$
$Word_0$		
⋮	c	d
$Word_i$	a	b
⋮		
$Word_m$	c	d

$$a = P(Word_i\ ,\ SPAM) \quad (2)$$

$$b = P(Word_i\ ,\ NOT\ SPAM) \quad (3)$$

連絡先：清水隆宏 e-mail:t.shimizu.tdu[at]gmail.com
東京電機大学大学院理工学研究科情報学専攻

$$c = \sum_{i \neq j} P(Word_j \ , \ SPAM) \quad (4)$$

$$d = \sum_{i \neq j} P(Word_j \ , \ NOT\ SPAM) \quad (5)$$

4. シミュレーション

本論のシミュレーションでは，はじめに教師情報として ham と spam の英文のメールデータを与える．教示するデータ数は，ham は 2500，spam は 500 とする．その後，学習を終えた後に easyham，hardham，spam の三種類のテストデータを与え，各分類器が正しく判別できるかテストを行う．テストデータの数は，easyham が 1200，hardham が 250，spam が 500 である．

hardham は教示データとして与えた spam 中の Word が含まれるため，easyham よりも判別が困難なテストデータである．

5. 結果

表 3，4 はシミュレーションの結果である．LS 分類器と Naive Bayes 分類器の結果を比較すると，LS 分類器は spam メールに対しての判別性能が Naive Bayes 分類器よりもすぐれていることがわかった．

しかし，easyham メールと hardham メールの判別では Naive Bayes 分類器よりも判別性能が悪く，hardham メールについては SPAM と判別する割合が大きいという結果となった．

表 3 ： LS 分類器による判別精度

	SPAM	*NOT SPAM*
easyham	*0.3778571*	*0.6221429*
hardham	*0.7500000*	*0.2500000*
spam	*0.8517192*	*0.1482808*

表 4 ： Naive Bayes 分類器による判別精度

	SPAM	*NOT SPAM*
easyham	*0.1742857*	*0.8257143*
hardham	*0.2338710*	*0.7661290*
spam	*0.5587393*	*0.4412607*

6. 考察

表 3 の結果から，LS 分類器は easyham，hardham，spam ともに SPAM と分類する傾向が見られる．特に，表 4 の Naive Bayes 分類器では hardham を 7 割以上で NOT SPAM と判別しているが，LS 分類器では 7 割以上で SPAM と判別している．

これは，本論の提案手法では LS を用いて各 Word の SPAM，NOT SPAM の判別をおこない，そして，メール本文中の SPAM，NOT SPAM の Word の割合でメール自体の判別を行うことが起因していると考えられる．

個々の Word において，LS は LS(SPAM｜Word)の計算を行う際に着目していない Word と NOT SPAM の情報を用いている．そのため，相互排他性バイアスの働きにより，着目していない Word と NOT SPAM の共起頻度が高まるにつれて LS(SPAM｜Word)の確信度も高まると考えられる．特に，本論では easyham のサンプル数が多いため，NOT SPAM の共起頻度が高まりやすい．

以上のことから各 LS(SPAM｜Word)の確信度が高まり，メール自体においても SPAM と判別する傾向が生まれると考えられる．

7. おわりに

本論では，Naive Bayes 分類器との判別性能を比較することで LS モデルがスパム分類タスクにおいて有用に働くことができるのか検証した．結果として，Naive Bayes 分類器を上回る精度は達成できなかった．しかし，今回提案した単純な手法においても高精度とはいかないがスパムメールの判別が可能であった．

今後の課題として，LS 分類器の SPAM 判別能力を維持したまま NOT SPAM の判別向上を目指す．

参考文献

[篠原 2007] 篠原修二，田口亮，桂田浩一，新田恒雄: 因果性に基づく信念形成モデルと N 本腕バンディット問題へ応用，人工知能学会論文誌 22 巻 1 号 G，pp.58-68，2007.

[篠原 2006] 篠原修二，田口亮，桂田浩一，新田恒雄:語彙獲得課題における非論理的推論の有効性，*ISPJ Symposium Series* Vol. 2006, No. 10, pp. 187-194，2008.

[甲野 2010] 甲野佑，高橋達二:因果帰納の調和対称ヒューリスティクス，日本認知科学会第 27 回大会発表論文集，43-46，2010.

[高橋 2010] 高橋達二，菅野宏明:因果帰納の対称性と非対称性，日本認知科学会第 27 回大会発表論文集，199-200，2010.

[Takahashi 2011] Tatsuji Takahashi, Shuji Shinohara, Kuratomo Oyo, Asaki Nishikawa, Cognitive Symmetries as Bases for Anticipation: a Model of Vygotskyan Development Applied to Word Learning, *International Journal of Computing Anticipatory Systems*, 24, 95-106, 2011.

[Conway 2012] Drew Conway, John Myles White, Machin Learning for Hacker, 2012.