3F4-2

スパム分類器への認知特性の付加とその分類の特徴

Addition of cognitive properties to a spam classifier

谷口 英貴[*1] 甲野 佑[*1] 清水 隆宏[*1] 高橋達二[*2]
Hidetaka Taniguchi Yuu Kouno Takayuki Shimizu Tatsuji Takahashi

[*1] 東京電機大学大学院 [*2] 東京電機大学
Graduate School Tokyo Denki University Tokyo Denki University

Previous studies have shown that some combo of human cognitive biases is effective in machine learning. The well used model of the biases is called LS (loosely symmetric) model. In this study, we test how LS works for classifying spam emails.

1. はじめに

先行研究では人間の因果関係の推論傾向を有するモデル（Loosely Symmetric model：LS）を通して，人間の認知特性が機械学習での諸タスクにおいて有効である事が示されている．

本研究ではより汎用的な機械学習タスクであるスパムメール分類を扱い，LS を通して人間の認知特性が既存モデルでは困難なメールの分類においてどの様に働くか検証した．

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].

人には，原因となる事象 p と結果となる事象 q がある時，”p→q”が真であれば”q→p”もまた真であると思い込む対称性バイアスがある．また”p→q”であるば”p→q”であると思い込む相互排他性バイアスも存在し，これは論理学においては裏と表の関係にあり，人は直感のみでは誤った解を導き出す恐れがあると言える．

しかし，LS は他の普遍性などを用いてこれらのバイアスを柔軟に変化することにより，人間の因果機能に対し高い相関を持ち，また，機械学習においても高い成果を出している．スパムメール分類器に人間の認知特性を組み込むことによるデータ判別の柔軟化が本研究に LS を採用する端緒となった． 本研究のモデルにおける a,b,c,d はそれぞれ p,q の共起頻度，あるいは共起確率 pq,p$\bar{q}$,$\bar{p}$q,$\bar{p}\bar{q}$に対応する．対応表を表 1 に示す．

表 1：共起情報の 2×2 分割表

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

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

連絡先：谷口英貴 e-mail : ht_msn[at]outlook.com
東京電機大学大学院理工学研究科情報学専攻

3. 教師あり学習

例えばスパムメールフィルタ等の，機械に未定義のデータを分類させる手法を機械学習という．機械学習は教師あり学習と教師なし学習の二種類に大別される．本研究で扱う教師あり学習は，既に与えられたサンプルデータとそれに付随される教師信号から機械的に解析し，その結果から教師信号に対する判別ルールを自律的に生成していく学習手法である．スパムメールフィルタは教師あり学習に分類され，ヘッダーや本文を含むメールデータと，迷惑メール（スパム）/非迷惑メール(ハム)という教師信号から，どのような単語や構造が含まれるとスパム/ハムに分類されるかを学習する．

しかし，ハムメールの中には easy ham と呼ばれる簡単にハムと分類できるメールに対して，hard ham と呼ばれるスパムとハムの境界にあるようなメールが存在し，これを正しく判別する事は困難である．本研究では教師あり学習(スパムフィルタ)に LS を組み込んだアルゴリズムを用いる事で，プログラムに人間に近い直感性を付与し，未知のデータに対してもより柔軟化かつ正確な判別を可能なプログラムの実装を目指す．

3.1 ナイーブベイズ 分類器

本研究におけるスパムフィルタはナイーブベイズ分類器と呼ばれるアルゴリズムを使用する．今回はメール本文中に使用されている単語のみを扱い，各単語のスパムメール内における使用頻度と，非スパムメール内における使用頻度とを調べる[Conway 2012]. これによりスパム・非スパムメールに含まれる任意の単語$word_i$の確率$P(word_i|class)$を学習し，このデータを元にメールが有害であるか無害であるかを判別する．式 2 を用いてメールの分類$(class)$がスパムである確率$P(spam|mail)$と非スパムである確率$P(ham|mail)$を計算し，その大小関係から任意のメール$(mail)$の分類を決定する．

$$P(class|mail) \propto P(mail|class)P(class)$$

$$P(mail|class) = \prod_{i}^{n} P(word_i|class) \quad (2)$$

3.2 第一種過誤と第二種過誤

既存モデルにおいては学習データ・実験データ双方共に存在しないデータに対する処理が曖昧であり，これが原因で第一種過誤と第二種過誤が発生する恐れがある．第一種過誤は疑陽性と呼ばれ，偽である情報を誤って真と判断してしまうことである．また，第二種過誤は偽陰性と呼ばれ，真である情報を偽と

判断してしまうことである. 本研究では人間の認知的性質を揺する LS を主観確率として用いる事で, スパムフィルタにおける二つの過誤の発生率にどのような変化が発生するか検証した.

4. LSNB 分類器

5 章で述べたナイーブベイズのように, 既存モデルを用いたスパムフィルタは未知のデータに対して適切な判断をし難い. そこでナイーブベイズに LS を組み込み, 教師情報を元に各単語のスパム・非スパムにおける共起情報を記録する. また, メールの判別を行うため教師情報を LS が扱える形に変換する. メール本文に存在する単語を共起頻度から抽出し, 任意の単語$word_i$とスパム・非スパムの共起情報を表 2 のように分割する. そして式 3 により, 通常のナイーブベイズで$P(class|mail)$を計算したように$LS(spam|mail)$と$LS(ham|mail)$を計算して, その大小関係によって任意のメール$(mail)$の分類を決定する.

表 2 : 抽出した共起頻度表

	$Spam$	Ham
$word_i$	a	b
$\neg word_i$	c	d

$$LS(class|mail) \propto LS(mail|class)P(class)$$

$$LS(mail|class) = \prod_{i}^{n} LS(word_i|class) \quad (3)$$

5. シミュレーション

始めに教師情報としてスパム・非スパムの英文メールデータをスパムフィルタに与え, それぞれデータ数は 2673 個, 492 個とした. スパムフィルタの学習を終えた後に, 非スパムであると判別が容易なメール(以下 easy ham), 非スパムであると判別が困難なメール(以下 hard ham), スパムメール(以下 spam)の三種類のデータを与え, 各分類機が正しく判別可能かテストを行った.

5.1 結果

LS を用いた分類器はナイーブベイズのシミュレーションの結果を表 3 に, ナイーブベイズのみの結果を表 4 に示す. この結果から LS を用いた分類器はナイーブベイズのみの分類機に比べ, spam の判別性能が向上していることがわかる. また, hard ham においても若干の向上が見られるが, easy ham の判別性能は低下している.

表 3 : LS 分類器による判別精度

	spam	Ham
easy ham	0.3657143	0.63428571
hard ham	0.2661290	0.73387097
Spam	0.9204871	0.07951289

表 4 : ナイーブベイズ分類器による判別精度

	spam	Ham
easy ham	0.1957143	0.8042857
hard ham	0.2782258	0.7217742
spam	0.8495702	0.1504298

5.2 考察

ナイーブベイズに LS を組み込むことで, spam の分類においても ham の教師情報を, ham の分類においても spam の教師情報を参照することが可能となり、相互排他的な分類が可能となった. これに対し, 従来のナイーブベイズは未知の単語が出現した際に, その単語がスパムである確率を低く見積もって計算するため, 正しい判別をすることがより困難であると考えられる.

しかし, LS を用いた判別器では既存の分類器よりも ham を spam として判別してしまった. これは Easy ham は, hard ham や spam の教師情報に対してデータ数が多く, spam メールに含まれる無害な単語を多く参照してしまったことが予想される. 以上のことから, easy ham と spam の分類において, スパムと判別される傾向が高まったと予想される.

6. 結論

本研究では, ナイーブベイズに LS を組み込んだ分類器と既存モデルとの比較を行い, LS がスパム分類器においてどのように働くかを検証した. spam 及び hard ham においては LS が有用に働くことを確認できたが, より多くの情報を含んだデータに対しては誤った判断を下す傾向が強くなってしまった. 今後の発展として, よりスパムの検出率を高めると共に, 非スパムメールの判別の向上を目指す.

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.

[清水 11] 清水 隆宏, 横川 純貴, 甲野 佑, 高橋 達二:認知バイアス調整機構 LS の Q 学習への実装とその機能, JSAI 2011(2011 年度人工知能学会全国大会(第 25 回)) , 予稿集, 2011

[清水 13] 清水 隆宏, 大用 庫智, 高橋 達二:人間の因果的直観を用いたスパム分類器, JSAI2013(2013 年度人工知能学会全国大会(第 27 回)) , 予稿集, 2013

[Takahashi 2010]. T. Takahashi, M. Nakano and S. Shinohara, *“Cognitive symmetry:
Illogical but rational biases,”* Symmetry: Culture and Science, Vol. 21, No. 1-3, pp. 275-294, 2010.

[Takahashi 2011]. T. Takahashi, K. Oyo and S. Shinohara, *“A Loosely Symmetric Model of Cognition,nitiakahashi, K. Oyo and S. Science, No. 5778, Springer, pp. 234-241, 2011.

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