6C-06

ベイズ情報量規準を用いた素性選択に基づく話題語検出

福重 貴雄† fukushige.yoshio@jp.panasonic.com
パナソニック株式会社

1 はじめに

テキストデータの時系列的変化の検知が重要となる応用分野として顧客からの相談分析がある．筆者は変化点検出問題を分類器における素性選択問題に帰着させ，ベイズ情報量規準を用いて相談データにおける出現確率が急増大した語を話題語として抽出する手法を開発した．本稿では開発した話題語抽出手法について報告する．

2 変化点検出のための従来手法と本稿の提案手法

本節では変化点検出のための主な従来手法[1]と本稿提案の方法との違いを述べる．以下で「対照期間」は比較の基準となる期間，「注目期間」は変化を検出したい期間を意味する．

2.1 外れ値検出を行う方法

対照期間のデータから，語の出現確率分布を推定し，推定された分布の下で，注目期間の出現データが実現する確率を計算し，一定の閾値以下になった場合に，出現データを「従来のデータからは起こり得ない」データとして，特異データとする方法である．

2.2 統計的分布の違いを検出する方法

対照期間のデータと，注目期間のデータからそれぞれ語の出現に関する確率分布を推定し，それらの間の分布としての差を評価した値が一定以上になった場合に，特異データとする方法である．χ二乗検定などの統計的検定のほか，カルバック・ライブラー密度比推定法などがある．

2.3 本稿の提案手法

本稿で提案する手法においては，注目する期間の問合せを特徴づける語を検出する．より具体的には，注目期間における問合せと対照期間における問合せの分類を行う分類器を教師あり学習により生成し，問い合わせデータにおける出現・非出現が同分類器において分類に有効な素性として使われるような語を話題語として検出する．

3 提案手法の詳細

3.1 基本的なアプローチ

注目期間と対照期間のデータにおける各語wについて，注目期間と対照期間で異なるパラメータにより生成されるモデル$\mathcal{M}_1$と，全期間に共通なパラメータにより生成されるモデルM_0とをベイズ情報量規準(BIC)に基づいて比較し，$\mathcal{M}_1$のほうが$\mathcal{M}_0$より優れており，かつ，$\mathcal{M}_1$において注目期間での出現確率が対照期間での出現確率より高いと判定された場合には，wは対照期間に特徴的な語（話題語）であるとし，そうでない場合は話題語でないとする．

Topic Words Detection based on Feature Selection using Bayesian Information Criterion

†Yoshio FUKUSHIGE,
Panasonic Corporation

3.2 生成モデル

注目期間中の問合せをクラス A,対照期間中の問合せをクラス B とし，各問合せが属すクラスを表す確率変数を c，語 w の当該問合せでの出現状態を表す確率変数を x_w(1なら出現, 0なら非出現）とする．c は，確率 π のベルヌーイ分布[2]に従い，π は，パラメータ α，β を持つベータ分布に従うとする．さらに，各語は A での分布と B での分布が異なる語の集合 W_1と，A での分布と B での分布が共通である語の集合 W_0のいずれかに属し，語 w が W_1に属す場合，A に属す問合せにおいて x_w は確率$\pi_w^{(A)}$を持つベルヌーイ分布に従い, B に属す問合せにおいて x_wは確率$\pi_w^{(B)}$を持つベルヌーイ分布に従うとする．語 w が W_0に属す場合，x_w は確率$\pi_w^{(0)}$を持つベルヌーイ分布に従うとする. さらに, $\pi_w^{(*)}$は, パラメータ$\alpha_w^{(*)}, \beta_w^{(*)}$を持つベータ分布に従うとする（*=A, B, 0）Figure 1は上記で仮定したモデルをグラフィカルに表現したものである[3].

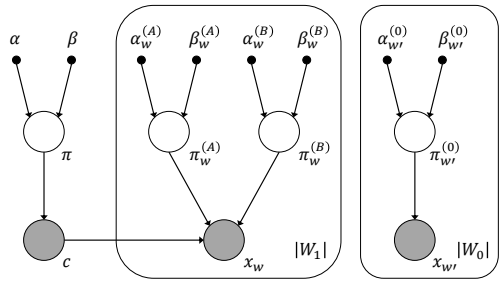

Figure 1

3.3 ベイズ情報量規準

ベイズ情報量規準 (Bayesian Information Criterion, BIC)[4]は，モデルの事後確率に-2をかけたものを次式により近似したもので，BIC を用いたモデル選択では，BIC が小さいモデルを選択することにより，事後確率が大きなモデルを選択する．ただし，$\hat{\boldsymbol{\theta}}$ はモデルにおけるパラメータの最尤推定量のベクトルで，kは，独立変数の数(=$\hat{\boldsymbol{\theta}}$ の次数)，Nは，データの個数である．

(1) $$\mathrm{BIC} \equiv -2\ln p\left(\mathcal{D}\middle|\hat{\boldsymbol{\theta}}\right) + \mathrm{k}\ln N$$

上記モデルの BIC は以下のように計算できる．

(2)
$$\begin{aligned}\mathrm{BIC} = &-2\sum_{c=A,B} N_c \ln p(c|\alpha,\beta)\\ &-2\sum_{w\in W_1}\sum_{c=A,B} N_{c,+w}\ln p\left(x_w=1\middle|\alpha_w^{(c)},\beta_w^{(c)}\right)\\ &-2\sum_{w\in W_1}\sum_{c=A,B} N_{c,-w}\ln p\left(x_w=0\middle|\alpha_w^{(c)},\beta_w^{(c)}\right)\\ &-2\sum_{w\in W_0} N_{+w}\ln p\left(x_w=1\middle|\alpha_w^{(0)},\beta_w^{(0)}\right)\\ &-2\sum_{w\in W_0} N_{-w}\ln p\left(x_w=0\middle|\alpha_w^{(0)},\beta_w^{(0)}\right)\\ &+(1+2|W_1|+|W_0|)\ln N\end{aligned}$$

ただし，$N_{c,+w}$（c=A,B）はクラス c の問合せで語 w が出現した問い合わせ数，$N_{c,-w}$ はクラス c の問合せで語 w が出現しなかった問い合わせ数，N_{+w} は語 w が出現した全問い合わせ数，N_{-w} は語 w が出現しなかった全問い合わせ数，N は全問い合わせ数である．

3.4 語の話題性の判定

$\mathcal{M}_1$ と $\mathcal{M}_0$ の BIC の差 −ΔBICは，以下のように計算される．

(3)
$$
\begin{aligned}
-\Delta\mathrm{BIC} &= -\mathrm{BIC}_{\mathcal{M}_1} + \mathrm{BIC}_{\mathcal{M}_0} \\
&= 2\sum_{c=A,B} N_{c,+w} \ln p\left(x_w = 1 \middle| \alpha_w^{(c)}, \beta_w^{(c)}\right) \\
&+2\sum_{c=A,B} N_{c,-w} \ln p\left(x_w = 0 \middle| \alpha_w^{(c)}, \beta_w^{(c)}\right) \\
&-2N_{+w} \ln p\left(x_w = 1 \middle| \alpha_w^{(0)}, \beta_w^{(0)}\right) \\
&-2N_{-w} \ln p\left(x_w = 0 \middle| \alpha_w^{(0)}, \beta_w^{(0)}\right) \\
&-\ln N
\end{aligned}
$$

−ΔBICが正でかつ，$\mathcal{M}_1$ において注目期間での出現確率が対照期間での出現確率より高いと判定された場合には，wは話題語であるとし，そうでない場合は話題語でないとする．また，-ΔBIC/N を話題語である度合いとする．

4 評価実験

4.1 対象データ

実験では，2018年4月1日～6月30日の各日について，当日を含む直近1週間を注目期間とし、話題語の検出を行った．対照期間は1週間空けて6週間前から2週間前までの4週間とした．対象の問合せは 3つの AV 商品カテゴリに関する全問合せとした．

4.2 従来手法との比較

従来手法として，二項分布モデルによる異常値検出と，分布の同一性に関する χ 二乗検定を行った．二項分布モデルによる異常値検出においては，各語の対照期間データから作成した二項分布モデルに基づいて，注目期間における当該語の出現問い合わせ数を予測し，実際の出現問い合わせ数を超える確率が1%以下になるような語を話題語とした．χ 二乗検定においては，注目期間における分布（出現数）と対照期間における分布が同一であるという帰無仮説が1%の有意確率で否定される語を話題語とした．

Table 1

商品カテゴリ		A	B	C
総相談件数		38,562	76,251	3,826
本手法	平均話題語数	**12.1**	**9.6**	**3.3**
	異なり話題語数	258	263	102
	-ΔBIC/N の最大値（記録日）	0.0361（4 月 23 日）	0.0038（6 月 18 日）	0.0227（4 月 24 日）
	上記における話題語	再編	マルチアングル	ショップチャンネル
二項分布モデルによる外れ値検出	平均話題語数	122.1	165.6	19.9
	異なり話題語数	2,450	3,439	494
χ 二乗検定（有意水準 1%）	平均話題語数	42.1	54.86	7.6
	異なり話題語数	1,068	1,457	239

4.3 結果の概要

Table 1 に実験結果の概要を示す．表中の「検出回数」は，一語以上話題語が検出された日の数，「平均話題語数」は，話題語が検出された場合に同時に検出された話題語の数，「異なり話題語数」は，全期間を通して検出された話題語の異なり数である．また、抽出例として Figure2 に商品カテゴリ A について，いつどのような話題語が抽出されたかを示す．

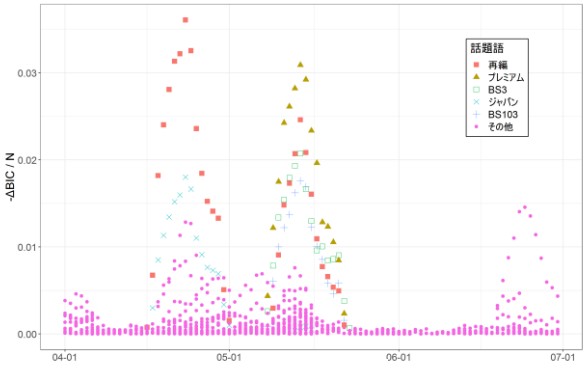

Figure 2

4.4 結果からの考察

実験期間には，4K8K 衛星放送の開始に向けた電波の再編が行われており，再編日前後には関連した問い合わせが急増した．結果を見ると関連したキーワードが話題語として適切に検出されていることがわかる．また，商品カテゴリ B の問合せにおける「マルチアングル」は同時期に人気アーティストのコンサートを収録したマルチアングル対応の DVD が発売された影響とみられる．商品カテゴリ C に関しては，4月中旬に問い合わせの急増が見られたが，抽出された話題語から，直前に通販番組で売り出されて，購入されたお客様からの問い合わせが原因であったことがわかった．一方，従来の二項分布にもとづく異常検知や χ 二乗検定では，非常に多くの語が抽出され，適切な結果が得られなかった．

5 まとめ

変化点検出問題を分類器における素性選択問題に帰着させ，ベイズ情報量規準を用いて出現確率が急増大した語を話題語として抽出する手法を開発した．お客様相談センターに寄せられた相談データを用いた評価実験では，従来の外れ値検出や統計的検定を用いた手法より優れた結果が得られることが確認できた．

6 参考文献

[1] 井出剛，杉間将：異常検知と変化検知，講談社，(2015).
[2] A.Gelman et.al.: Bayesian Data Analysis, Chapman & Hall, London, (1995).
[3] C.M.Bishop （元田浩 他訳）：パターン認識と機械学習，丸善出版(2012).
[4] 渡辺澄夫：ベイズ統計の理論と方法，コロナ社(2012).