E-008

トピックモデルを用いたソーシャルメディアからの市場シェア予測

A topic model for predicting market share from social media

中島 光夫 *　津川 翔 *　山本 幹雄 *
Mitsuo NAKAJIMA　Sho TSUGAWA　Mikio YAMAMOTO

1. はじめに

ソーシャルメディアでやりとりされている膨大なテキストの中には社会に対する意見や感情が含まれていると考えられており、このような情報から株価の上下や病気の流行などの現象を説明・予測する研究が盛んに行われている。特に最近では Twitter などの SNS の予測力が注目されており、さまざまな予測対象に応用が広がっている。研究テーマとしては、対象変数の予測に役立つ素性をいかに抽出するかが盛んに研究されている [Bollen et al., 2011][Si et al., 2013]。

本稿では、Twitter のデータから複数の競合商品の売上げ (順位) や売上げ比率 (シェア) を予測する問題を検討する。素性としては、supervised-LDA(以下 sLDA)[D. Blei and J. McAuliffe, 2007] の考えを踏襲し、トピックモデルが出力するトピック比率が文書集合の要約になっていると考える。さらに、sLDA では一次元の値を出力するモデルであった点を拡張し、多次元の値 (シェア) を同時に予測するモデルを検討する。また実際に車種ごとの自動車売上の市場シェアを予測し、本稿の拡張は自動車売上の市場シェアを予測するのに有効であることを示す。

2.supervised LDA の拡張による市場シェア予測

文書集合から得られる特徴量のひとつとして、トピック比率がある。トピックモデルを用いて推定したトピック比率と回帰モデルを用いることで、文書から数値を予測したり、説明することができる。

sLDA は文書と値の対を学習し、入力された文書に対する値を予測するモデルである。予測を行う際は文書のトピック比率を推定し、その線形和を予測値とする。しかし sLDA は多次元の値を学習することができず、複数の文書から値を予測することもできない。そこで時刻 t の文書集合 D_t とそれに対応するベクトル $\mathbf{y}_t$ の対の集合である $\{(D_1, \mathbf{y}_1), \cdots, (D_T, \mathbf{y}_T)\}$ を学習・予測するようにモデルを拡張する。sLDA を拡張したモデルのグラフィカルモデルを図 1 に、諸元を表 1 に示す。$\boldsymbol{\theta}_d$ と $\boldsymbol{\phi}_k$ の確率はそれぞれパラメータ α、β のディレクレ分布に従い、$w_{d,n}$ の確率はパラメータ $\boldsymbol{\phi}_{z_{d,n}}$ の多項分布に従うと仮定している。このモデルにおいて Z、W、Y、Θ、Ψ が得られる確率は次のように表される。

$$
\begin{aligned}
&p(\mathsf{Z}, \mathsf{W}, \mathsf{Y}, \Theta, \Psi; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathsf{B}, \sigma^2) \\
=&\prod_{k=1}^{K} p(\boldsymbol{\phi}_k; \boldsymbol{\beta}) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d; \boldsymbol{\alpha}) \\
&\times \prod_{n=1}^{N} p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|\boldsymbol{\phi}_{z_{d,n}}) \prod_{t=1}^{T} p(\mathbf{y}|\overline{z}_t, \mathsf{B}, \sigma^2)
\end{aligned}
$$

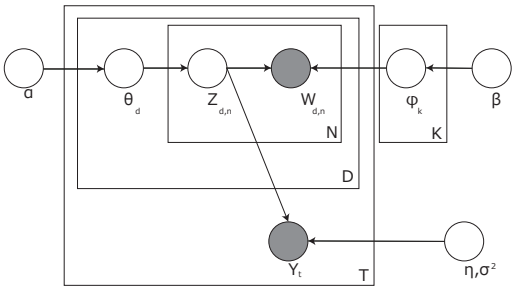

図 1　本稿で提案する拡張のグラフィカルモデル

表 1　諸元

D_t	時刻 t の文書集合
$w_{d,n}$	文書 d の n 番目の単語
$\mathbf{y}_t$	時刻 t の従属変数の値
T	総期間数
K	トピック数
L	従属変数のベクトルの次元
σ^2	従属変数の分散
α,β	トピック/単語の事前分布の超パラメータ
$z_{d,n}$	$w_{d,n}$ に対するトピックの割り当て
$\boldsymbol{\theta}_d$	文書 d の単語分布のパラメータ
ψ_k	トピック k の単語分布のパラメータ
$\overline{z}_t$	時刻 t における各文書のトピック比率の平均
$\mathbf{b}_l$	$y_{t,l}$ を予測するための回帰係数

便宜上 $w_{d,n}$、$z_{d,n}$、$\boldsymbol{\theta}_d$、ψ_k、$\mathbf{b}_l$ の行列表現をそれぞれ W、Z、Θ、Ψ、B と表した。これを変形して次の Gibbs Sampler を得る。

$$
\begin{aligned}
&p(z_{d,n}|\mathbf{Z}_{\neg(d,n)}, \mathbf{W}, \mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{B}, \sigma^2) \\
\propto&(n_{d,(\cdot)}^{k,\neg(d,n)} + \boldsymbol{\alpha}_k) \frac{n_{d,(\cdot)}^{k,\neg(d,n)} + \boldsymbol{\alpha}_k}{\sum_{v' \in V} n_{(\cdot),v'}^{k} + \boldsymbol{\beta}_{v'}} \\
&\times \prod_{l=1}^{L} \exp\left(\frac{b_{l,k}}{N_t \sigma^2} \left(y_{t,l} - \mathbf{b}_l^{\top} \overline{\mathbf{z}}_t^{\neg n} - \frac{b_{l,k}}{2N_t} \right) \right)
\end{aligned}
$$

ここで $\mathbf{Z}_{\neg(d,n)}$ は $z_{(d,n)}$ 以外の全てのトピックの割り当てを表し、$n_{d,(\cdot)}^{k,\neg(d,n)}$ は文書 d のうち n 番目を除く全単語のうちでトピック k に割り当てられたものの数を表す。これを全ての t、d、k、l について繰り返しサンプリングすることで、Markov chain Monte Carlo 法の枠組みで Z、Θ、Ψ を推定することができる (E-step)。この結果を用いて、予測誤差を最小にする B を決定する (M-step)。この E-step と M-step を交互に繰り返すことにより、最終的なパラメータの推定値を得る。

新しいドキュメント集合 D_t が入力された時、それに対応する数値の予測値 $\hat{\mathbf{y}}_t$ は $\hat{\mathbf{y}}_t = \mathsf{B}\overline{z}_t$ で与えられる。

* 筑波大学大学院システム情報工学研究科 mnakajima@mibel.cs.tsukuba.ac.jp

なお、このモデルでは値の分散は出力の各次元に関して共通であるが、次元ごとに個別に与えられるモデルへ容易に拡張することができる。

3. 自動車の市場シェア予測による評価実験

Twitter のデータを用いて競合する製品の市場シェアをどの程度推定できるか評価する。Twitter API の sample stream から得られた約 20 億件の投稿を用い、2010 年 3 月 1 日から 2012 年 12 月 31 日までの 2 年分を学習区間とし、それ以降 2013 年 7 月 31 日までの 7 ヶ月をテスト区間とする。予測する対象は新車乗用車販売台数月別ランキング[1]の 2010 年 4 月における上位 12 車種[2] を用いる。なお販売台数を全体に占める割合（市場シェア）に変換する。

Twitter から得られたコーパスに対して、自動車に関係ない投稿を除外し、高速に処理できるようにするため次の前処理を施す。これにより最終的に 954,042 件の投稿を得た。

1. 次のいずれかに当てはまる投稿を抽出
   - 日本語または英語の車種名を含む
   - いずれかの社名を含む
   - 自動車、車、くるま、クルマ、car、運転、ドライブ、ドライバーのいずれかの単語を含む
2. 抽出された投稿を形態素解析し、単語分割
3. 全ての文書において、全期間中の出現回数の上位 1 万単語以外を削除

比較のため 1. 車種名の月ごとの出現回数を特徴量とした重回帰モデル（正則化なし）2. 車種名の月ごとの出現回数を特徴量とした L_1 正則化項つきの重回帰モデル (LASSO 回帰) 3. 各車種それぞれに対して個別に sLDA を用いる方法による予測も行う。LASSO 回帰の正則化パラメータは分割交差検定を用いて決定した。sLDA のパラメータは、本稿の拡張と同じ設定を用いた。

推定したシェアの順位が実際の順位とどの程度一致するかに注目するため、評価指標として Kendall の順位相関係数を用いる。これは 2 つの順位がどの程度一致しているかを測る指標で、1 から −1 までの値をとる。1 に近いほど両者は似た順位となり、−1 に近いほど両者は異なる順位となる。この値が 1 の場合は完全に一致し、−1 の場合は真逆の順位となる。

訓練区間、テスト区間それぞれにおける順位相関係数の平均値を表 2 に示す[3] 。また年月を横軸にとった順位相関係数のグラフを図 2 に示す。

sLDA と提案するモデルでは、提案するモデルが上回っている。12 車種全体の説明変数を同時に推定することで、12 車種それぞれに対して個別に sLDA を用いて説明変数（トピック）を推定する場合よりも未知の入力に対して頑健な特徴量が抽出されたためではないかと考えられる。LASSO 回帰による予測がテスト区間平均における最高値の 0.610 を示したが、提案する拡張も同程度の 0.602 となった。一方、正則化を行わない重回帰モデルは、訓練区間においては高い性能を示すにもかかわらず、テスト区間では最低の性能となっている。これは訓練データに対して過適応を起こしているためである。

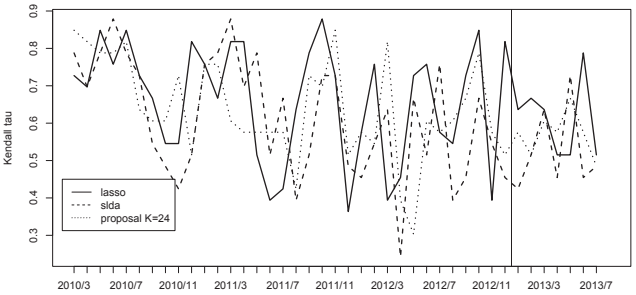

図 2　順位相関係数の時系列

表 2　順位相関係数のテスト区間平均と訓練区間平均

model	テスト区間平均	訓練区間平均
重回帰	0.190	0.660
LASSO	0.610	0.663
sLDA	0.528	0.615
proposal($K = 24$)	0.602	0.638

4. おわりに

Twitter のデータを用いて市場シェアの予測するため、トピックモデルのひとつである sLDA を拡張した。実験では自動車の市場シェアの予測を行い、予測したシェアと実際のシェアの間には中程度の相関があることが示された。

今後、sLDA 内で用いる回帰モデルとして、非線形なモデルを含むさまざまなモデルを試すことで、性能の向上が期待できる。また、自動車のシェア以外の問題に対してどの程度有用であるかを検証する必要がある。

参考文献

[Bollen et al., 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[D. Blei and J. McAuliffe, 2007] D. Blei and J. McAuliffe (2007). Supervised topic models. In *Neural Information Processing Systems*, 21.

[Si et al., 2013] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (2011):24–29.

[1] `http://www.jada.or.jp/contents/data/ranking.html` （日本自動車販売協会連合会）

[2] プリウス/フィット /ヴィッツ/カローラ/パッソ /セレナ/ステップワゴン/フリード/デミオ/ヴォクシー/ノア/スイフト

[3] 各種パラメータは $K = 24$、$\alpha = 1.0$、$\beta = 1.0$、$\sigma^2 = 1.0$、とした。サンプリングの繰り返しは 40 回、E-step、M-step の実行回数は 40 回に固定した。