様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　１２１０２
研究種目：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１２１４９

研究課題名（和文）テキストデータに対する高次元小標本回帰問題へのトピックモデルに基づくアプローチ

研究課題名（英文）An aproach to high dimensional regression problems on small text data using topic models

研究代表者

山本　幹雄（YAMAMOTO, Mikio）

筑波大学・システム情報系・教授

研究者番号：４０２１０５６２

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では、(1)トピックモデルを用いた少量のテキストデータからの高次元回帰問題と、(2)web上のユーザの検索行動データを状態空間モデルに統合した自動車販売台数予測問題に対する検討を行った。成果は以下である。
(1) リッジ回帰やlasso回帰などのいくつかの縮小推定手法が、教師ありトピックモデルの性能を改善するために効果的であることを実験的に示した。
(2) 検索行動量が状態空間モデルを利用した自動車販売台数予測問題の精度を上げるために使えることを実験的に示した。

研究成果の概要（英文）：In this research, we investigated (1) high dimensional regression problems using topic models for small text data and (2) prediction problems of car sales using state space models with the search behavior of users on the web. The achievements are the followings.
(1) We showed that various shrinkage estimation methods such as ridge and lasso regressions are effective in order to improve supervised topic models for high dimensional and small text data.
(2) We showed that the search behavior volume data can be used for increasing the accuracy of car sales prediction using state space models.

研究分野：　情報工学

キーワード：　トピックモデル　縮小推定　検索行動量　状態空間モデル

１．研究開始当初の背景

近年、ユーザが入力・発信したweb上の大規模な情報を元に、インフルエンザの流行や経済指標などのマクロな動向・時系列を推定・予測する手法の研究が盛んである①。これは、ソーシャルメディアの発達と普及によって、リアルタイムに全世界の人々の感想や意見を収集・利用できるようになったことによる。直接観測できない全世界の人々の意見や動向を集約する手法の研究は様々な分野で期待されており、今後ますます重要になると考える。

しかし、テキストデータから得られる特徴ベクトルは、単語をベースとすると単語種類数の次元となり、非常に高次元(数千～数万次元) かつスパースである。一方、予測の対象となる数値データの数が多くない場合、典型的な「高次元小標本」問題となり、妥当なモデルの推定は非常に挑戦的な課題となる。

また、web上にはテキストデータ以外にも予測に利用できるデータは多い。どのようなデータを利用すれば予測精度を改善できるかの事例的な研究を積み重ねることが重要である。

２．研究の目的

本研究では、ソーシャル・メディア上に一般ユーザが日々公開しているテキストデータやWeb上の行動から、商品の販売数や評価値などの数値を予測することを目的とする。本研究では、入力データがテキストである点と、典型的な高次元小標本からの回帰問題に分類される点に着目した手法を開発する。具体的には、高次元小標本回帰問題に対する代表的手法であるLassoなどの縮小推定手法をベースに、これらが苦手とするテキストデータの性質をトピックモデルを組み合わせて解消する手法を開発する。また、扱う対象が時系列であるため時系列の汎用的なモデル化手法である状態空間モデルにWeb上のデータを加える統合手法を検討し、予測精度を高める手法を開発する。

３．研究の方法

以下、本研究で開発したテキストデータに対するトピックモデルと縮小推定を用いた回帰の組み合わせ手法を(1)で、Web 上での行動量を時系列予測モデルに組み込む手法を(2)で説明する。

（1）提案モデルの中心的アイデアはL_1正則化等の縮小推定手法が苦手とするテキストデータの性質(スパース性と高相関次元の存在)を、トピックモデルによる縮約により回避する点にある。具体的には、supervised LDA②（以下、sLDAと略記する）を基本とし、sLDA の線形回帰部分をL_1正則化等の様々な縮小推定手法を組み合わせたときの性能を実験的に調査することにより、組み合わせの良否および効果的な組み合わせを特定した。

sLDAは次のような生成過程でモデル化される②。

1．トピック比率 $\theta \sim \mathrm{Dir}(\alpha)$
2．文書中の各単語w_nについて:
(a)トピック割当て $z_n \sim \mathrm{Mult}(\theta)$
(b)単語 $w_n \sim \mathrm{Mult}(\beta_{zn})$
3.予測値 $y \sim N(\mu, \sigma^2)$

ここで、$\mathrm{Dir}(\alpha)$はトピック数Dの次元を持つDirichlet分布、$\mathrm{Mult}(\theta)$および $\mathrm{Mult}(\beta_{zn})$は$\theta$または$\beta_{zn}$を確率とした場合の多項分布、$N(\mu, \sigma^2)$は平均$\mu$,分散$\sigma^2$の正規分布。ただし、$\mu$は文書中のzの平均値を説明変数として回帰した場合の結果であり、以下のように定義される。

$$\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n, \quad \mu = \eta^\top \bar{z}$$

ηは回帰係数である。本研究ではこの回帰係数ηを推定する際に様々な縮小推定を利用し、小標本問題による性能悪化を改善する。

M個の文書から成る文書集合$D = \{d_1, d_2, \ldots d_M\}$と、それぞれの文書が示す数値$Y = \{y_1, y_2, \ldots, y_M\}$が与えられたとき、パラメータを推定するために本研究ではcollapsed Gibbs samplingを用いたMCMC法とEMアルゴリズムを組み合わせた手法を用いた。次のような2ステップを収束するまで繰り返す。

(step1) 次のcollapsed Gibbs Samplingによって単語のトピック割当て$z_{d,n}$をサンプリングする③。

$$p(z_{d,n}=k|w, y_d, \eta, z_{-(d,n)}) \propto (n_{d,k}^{-d,n}+\alpha)\frac{n_{w_{d,n},k}^{-d,n}+\zeta}{N_k^{-d,n}+W\zeta} \cdot \exp\left(2\frac{\eta_k}{N_d}(y_d - \eta^\top \bar{z}_d^{-n}) - \left(\frac{\eta_k}{N_d}\right)^2\right)$$

ここで、"d,n"の添字は文書dのn番目の単語を意味する。"-d,n"の添字は文書dのn番目の単語を除いた場合を意味する。nは指定された単語の数、Nは総数である。例えば、$n_{d,k}^{-d,n}$は、文書dのn番目の単語を除いた場合の文書d中でkというトピック割当てを持つ単語の数である。Wは単語の種類数、αとζはスカラのハイパーパラメータである。

(step2) サンプリングされたトピック割当ての割合$\bar{z}_d$からy_dを予測する回帰式の重みηを推定する。推定手法として、縮小推定ではない二乗誤差最小基準による方法と、縮小推定法として、リッジ回帰、Lasso、Elastic-net、主成分回帰を用いた④。成果の章でそれぞれの推定手法を用いた場合を比較する。主成分回帰以外の最適化基準（誤差関数）は以下の一般式で説明できる。

$$Err = \sum_d (y_d - \eta^\top \bar{z}_d)^2 + \lambda_1 \sum_k \eta_k^2 + \lambda_2 \sum_k |\eta_k|$$

最小二乗法は$\lambda_1=\lambda_2=0$、リッジ回帰は$\lambda_2=0$, Lassoは$\lambda_1=0$、Elastic-netはλ_1とλ_2の両方がゼロでない場合である。*Err*を最小とするηを求める。主成分回帰は$\bar{z}_d$の主成分を用いて回帰を行う。

(2) 検索行動量を用いた時系列推定手法を開発した。自動車等の比較的高額な商品の購入において、消費者は特に念入りな事前調査を行うであろうという仮説を置き、事前調査の総量を反映する量を利用して新車販売台数予測の精度を改善する手法を検討した。特に、事前調査は購入の前に行われるため、未来の販売数を予測する場合に効果的であると考えた。事前調査の総量を反映する具体的な量としては、直接的な検索数を反映するGoogle Trendsの時系列、また、実際の調査対象となるページへのアクセス数を反映するWikipedia閲覧数の時系列を用いる。

基本的な時系列予測には状態空間モデルによる次の式を用いる。tは時刻である。

$$y_t = \mu_t + s_t + u_t, \quad u_t \sim N(0, U)$$
$$\mu_t = 2\mu_{t-1} - \mu_{t-2} + v_t, \quad v_t \sim N(0, V)$$
$$s_t = -\sum_{l=1}^{11} s_{t-l} + w_t, \quad w_t \sim N(0, W)$$

y_tが販売台数などの予測したい変数でありy_tの式が観測方程式である。μ_tとs_tが状態方程式であり、μ_tがトレンド成分、s_tが季節成分をモデル化する。この基本式をベースラインとし、これに検索行動量を組み込むことによって予測性能を改善する。検索行動量としては、対象商品をキーワードとした検索量（Google Trendsを用いる）、あるいはWikipediaの対象商品のページへのアクセス数を用いる。それぞれの行動量の時系列を販売数と同様にトレンド成分、季節成分に分解し、時間tをずらしたトレンド成分を基本式に加える。以下は行動量のモデルである。

$$e_t = \mu_t^e + s_t^e + u_t^e, \quad u_t^e \sim N(0, U^e)$$
$$\mu_t^e = 2\mu_{t-1}^e - \mu_{t-2}^e + v_t^e, \quad v_t^e \sim N(0, V^e)$$
$$s_t^e = -\sum_{l=1}^{11} s_{t-l}^e + w_t^e, \quad w_t^e \sim N(0, W^e)$$

検索行動量の時間をずらしたトレンド成分を最初の販売台数y_tの予測式に組み込む。トレンド成分の重みβも時変とする。

$$y_t = \mu_t + s_t + \beta_t \mu_{t-n}^e + u_t, \quad u_t \sim N(0, U)$$
$$\beta_t = \beta_{t-1} + v_\beta, \quad v_\beta \sim N(0, V_\beta)$$

nは固定で、時刻tにおけるy_tを予測するためにnヶ月前の検索行動量のトレンド成分を用いることを意味する。

４．研究成果

(1) トピックモデル(sLDA)と縮小推定の組み合わせの実験結果を述べる。データとしては、当初、マイクロブログのクローリングデータから自動車販売台数を予測対象として実施する予定であったが、マイクロブログのノイズが非常に多く有効な検証が行えなかったため、商品の評価サイトからクローリングした評価テキストと人間が付与した５段階評価の値を予測対象とした実験に切り替えた。ちなみに、マイクロブログのノイズは、商品名の曖昧さとととともにボットと呼ばれる自動生成された大量の投稿や、販売会社の販促キャンペーンによる投稿などが原因である。マイクロブログを用いた実験には、ノイズ除去手法の研究が別途必要である。

収集したデータはトレーニングデータとして1000、1万、10万件の３種類、開発データおよびテストデータとしてそれぞれ 1000 件を準備した。評価指標としては、1000 件のテストデータに対する予測数値と人間が付与した評価値との間の相関係数とした。

図 1～3 に、sLDA のトピック数を 2 から 5 まで変化させた場合のテストデータに対する相関係数を示す。各図は訓練データ量に対応する。エラーバーは標準偏差を示す。それぞれの縮小推定手法のパラメータは開発データで調整を行った。

全体的に、訓練データが 1000 件と少ない場合は、トピック数が 2 のときに最高性能となり、訓練データを増やしたときはトピック数が 3 のときに最高性能となることが分かる。また、訓練データ数が少ない場合は、比較的単純な縮小推定法であるリッジ回帰で十分な性能が出ているが、

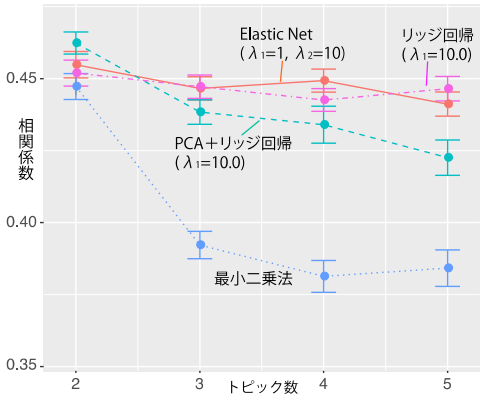

図 1　訓練データ 1000 件の場合

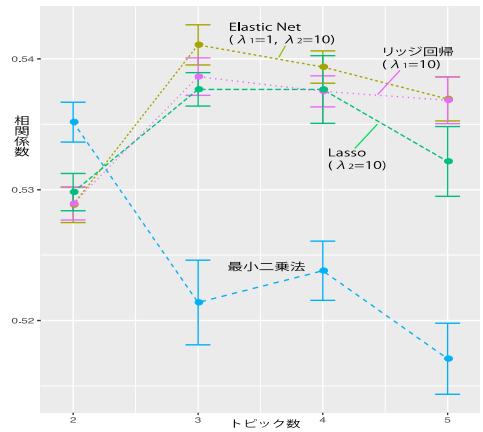

図 2　訓練データ 1 万件の場合

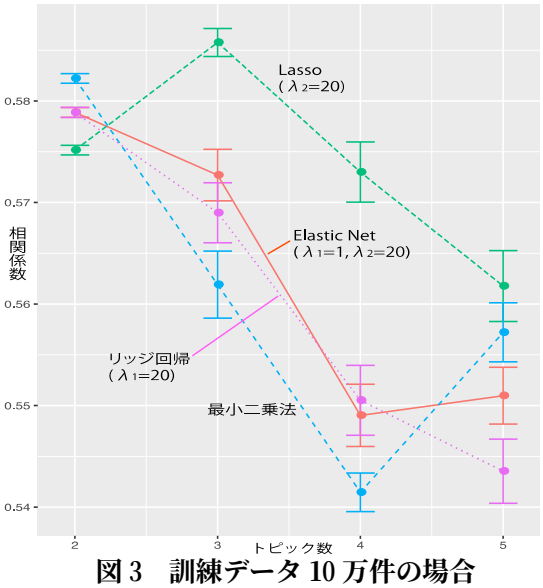

図 3　訓練データ 10 万件の場合

訓練データ数が増えた場合、Elastic Net や Lasso などのやや複雑な手法が高性能であることが分かった。訓練データ数が少ない場合は、トピック数が 2 という最小の場合で最高性能が出ているため、縮小推定の性能差が十分に発揮されていないと考える。

（2）次に、検索行動量を用いた時系列予測の精度改善に関する実験結果を示す。

実験データとして、1 ヶ月毎の人気車種 22 種類について新車登録台数を 2010 年 1 月から 2014 年 12 月までの 5 年間のデータを用いた。それぞれの製造会社と車名を検索キーワードとしてときに Google Trends と Wikipedia のページ閲覧数のデータを同じ期間集めた。前半の 4 年分を訓練データとし、最後の 1 年分をテストデータとして予測実験を行った。評価指標としては、予測台数と実際の新車登録数の二乗誤差の平方（RMS）を用いた。ただし、22 車種を平均する場合、絶対的な販売量が影響しないように相対的な RMS 値とした。

テストデータにある月の販売台数の具体的な予測方法は次の通りである。前半 4 年分のデータで最尤推定により状態空間方程式の分散（3 節(2)で提案したモデルの u_t, v_t, w_t など）を学習する。テストデータ(2014 年)の m 月の販売台数を n ヶ月前に予測する場合、

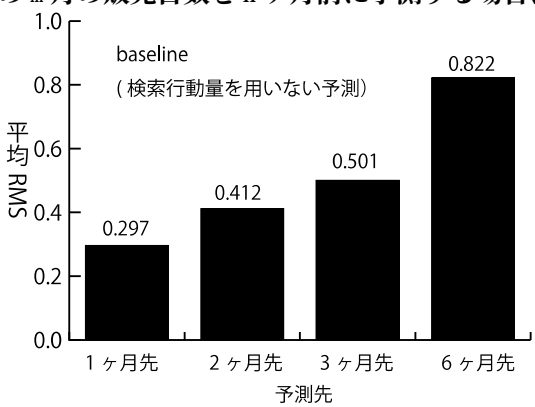

図 4　検索行動量を用いない場合の予測結果

m-n 月までの状態変数をカルマンフィルタによって計算し、そこから n ヶ月後である m 月の販売台数を予測する。

baseline として検索行動量を統合しない状態方程式による 1 ヶ月から 6 ヶ月先の販売台数を予測した場合の RMS 値（小さいほど正確）を図 4 に示す。当然ではあるが、近い未来を予測する方が簡単であることが分かる。

検索行動量を統合した場合の 1 ヶ月,2 ヶ月,3 ヶ月先の予測結果をそれぞれ図 5(a)～(c)に示す。統合する検索行動量は Google Trends と wikipedia のページ閲覧数であり、それぞれ何ヶ月前の検索量を統合したかを横軸で区別している。すべての場合で、検索行動量を統合することによって予測性能が改善されていることが分かる。また、Wikipedia のページ閲覧数よりも Google Trends を用いた方がよい改善につながっていることも分かる。さらに、場合によっては 3 ヶ月前の検索行動量を統合する場合が最も性能向上を達成できていることもあることが分かる。これは、ユーザがインターネットで情報を調べる行動した後、実際の新車登録されるまでには数ヶ月の時間遅れがあるからと説明できる。

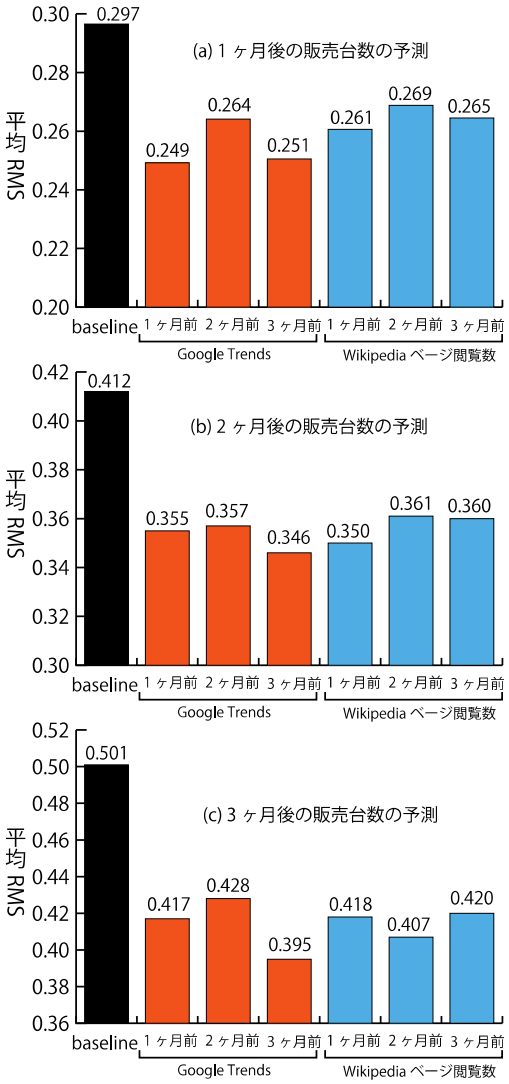

図 5 検索行動量を統合した場合の予測結果

＜引用文献＞

① 荒牧，増川，森田．「Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測」，情報処理学会研究報告，2011-NL-201(1)，pp.1-8，2011.

② D.M.Blei and J.D.McAuliffe. Supervised topic models, Advances in Neural Information Processing Systems 20, pp.121-128, 2007.

③ Jonathan Chang. Uncovering, understanding, and predicting links. PhD.Thesis. Princeton University, 2011.

④ T.Hastie, R.Tibshirani and J.Friedman. The elements of statistical learning. 2nd Ed. Springer, 2009.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計３件）

①山口　太一，角田　孝昭，吉田　光男，津川　翔，山本　幹雄．2017．検索行動量を用いた状態空間モデルによる自動車販売台数の予測 ．CQ 基礎講座ワークショップ，1page，January 2017.大阪大学中之島センター（大阪府大阪市）

②山口　太一，角田　孝昭，津川　翔，山本　幹雄．検索行動量を用いた自動車販売台数予測に必要な学習期間の長さについての分析．電子情報通信学会総合大会講演論文集(2016)，page 27．March 2016．九州大学（福岡県福岡市）

③角田　孝昭，吉田　光男，津川　翔，山本　幹雄．状態空間モデルを用いた検索トレンドとページビューからの自動車販売台数の予測．第29回人工知能学会全国大会論文集，4 pages，June 2015．はこだて未来大学（北海道函館市）

６．研究組織

(1)研究代表者

山本　幹雄（YAMAMOTO, Mikio）

筑波大学・システム情報系・教授

研究者番号：40210562

(4)研究協力者

角田　孝昭（TSUNODA, Takaaki）

山口　太一（YAMAGUCHI, Taichi）