トピックモデルに基づく気象要因と商品販売量の関係分析モデル

1X15C114-0　山之内 薫
指導教員　後藤 正幸

1　研究背景と目的

スーパーマーケットなどの小売店では，需要変動に起因する在庫過多や品切れに伴う機会損失の発生を回避するため，商品の販売量を販売履歴データから予測することが望まれている．こうした需要変動は様々な外的要因に影響されるが，本研究で対象とする生鮮食品も扱う小売店では，日々の生活と密接な関係にある天候や，行動パターンを決める曜日などの影響が特に大きいと考えられる．また，曜日や天候によって需要が大きく増減する商品がある場合には，品切れや売れ残りに直結し，大きな損失に結びついてしまう．すなわち，小売店の現場レベルでは，需要量のそのままの予測よりも需要の大幅な増減を検出することが望まれている．そこで本研究では，曜日と天候情報を入力とし，出力を需要の「大幅増」,「定常」,「大幅減」のカテゴリとする分類器による予測モデル構築を考える．

ここで，需要増減の分類問題に適用する際，一般的に分類器への入力は商品を One Hot ベクトルとして扱うことが考えられる．しかし，商品数が次元になるため，高次元かつスパースなベクトルとなり，実際のデータ量では学習がうまく行われない．ここで，文書やアイテムのような高次元なデータを潜在的な意味空間上の点として表現することができるトピックモデルの 1 つとして，Latent Dirichlet Allocation[1](以下，LDA) が知られている．LDA を用いることでデータの特徴量の次元を圧縮することができ，スパース性を解消できる．LDA を販売履歴データに適用させることにより，商品や日付に紐づく様々な要因による販売傾向を確率的に表現可能となる．加えて，商品ごとのトピック分布がわかるため，トピックについての解釈も可能になる．

以上により本研究では，LDA によって得られるトピック分布をを最大限に活用し，蓄積された販売履歴データに加え，天候や曜日データを考慮したもとで，その商品の販売量を「大幅増」,「定常」,「大幅減」のいずれかに分類するモデルを構築する．また，実店舗における販売履歴データを提案モデルに適用した分類実験を行い，提案モデルの有用性について評価・考察する．

2　提案モデル

本研究では，販売履歴データに LDA を適用して商品ごとのトピックへの所属確率分布を求め，販売量に大きく関わると考えられる天候と曜日の One Hot ベクトルとともに分類器への入力とするモデルを提案する．

2.1　対象とする天候要因

天候データは，気象庁から 1 時間ごとの降雨量，日射量および各月の晴天確率が発表される．しかし，需要変動には大まかな晴や雨などが関係すると考えられる．そこで各日の晴・曇・雨を以下のように定義する．天候を量的変数ではなく質的変数にすることで，対象を短期間にしているため，データ数が少なくとも推定精度が保つことができる．また，多少の降雨量の変化での需要の増減を検出したいわけではないため，晴・曇・雨とカテゴリで定義する．ここで，晴天確率とは各月の晴の日の確率である．

雨：開店時間帯のうち 1.0mm 以上の降水量が観測された時間が 5 時間以上あった日

晴：雨以外の日で，日射量が以下 (1) 式の境界値 R_m 以上であった日

$$R_m = \overline{S}_m \times (1 - P(c_m)) \tag{1}$$

曇：雨でも晴でもない場合

ここで，$\overline{S}_m$ は m 月の平均日射量，$P(c_m)$ は m 月の晴天確率である．

2.2　Latent Dirichlet Allocation

販売履歴データにおいて商品集合を $\mathcal{G}$=$\{g_1, \cdots, g_i, \cdots, g_I\}$，販売日の集合を $\mathcal{D} = \{d_1, \cdots, d_j, \cdots, d_J\}$，トピック集合を $\mathcal{Z}$=$\{z_1, \cdots, z_k, \cdots, z_K\}$ とする．また，販売日 d_j がトピック z_k に所属する確率を θ_{jk}，トピック z_k のもとで商品 g_i が出現する確率を ϕ_{ki} とする．このとき，トピック分布を $\boldsymbol{\theta}_j = (\theta_{j1}, \cdots, \theta_{jK})$，トピック z_k のもとで商品 g_i が出現する確率分布 (商品分布) を $\boldsymbol{\phi}_k$=$(\phi_{k1}, \cdots, \phi_{kI})$ と表す．ここで，$\boldsymbol{\theta}_j$，$\boldsymbol{\phi}_k$ にはそれぞれ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ をパラメータとするディリクレ事前分布を仮定する．販売日 d に商品 g が購入される確率は，

$$P(g|d)= \sum_{k=1}^{K} \int \int \theta_{jk} P(\theta_{jk}|\alpha_k) \phi_{ki} P(\phi_{ki}|\beta_i) d\theta_{jk} d\phi_{ki} \tag{2}$$

と表現される．ここで，販売日 d_j に販売された商品を g_{ji}，この商品に対応するトピックを z_{jk}，N_j を d_j に販売された総商品点数とする．このとき，グラフィカルモデルは図 1 のように表される．

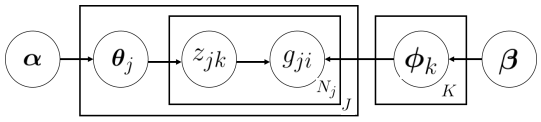

図 1．LDA のグラフィカルモデル

2.3　分類器

本研究の目的は，曜日や天候に起因する商品需要の大幅な変化を捉えることである．これは極めて多数の商品を扱っている小売店の現場では一定以上の変化が見込まれる場合に供給や調達の意思決定を変えるようなオペレーションが現実的である．そこで，大幅な需要増減の境界値として，商品ごとの販売数の中央値の ±30 %以上の増減を分類対象とし，以下のように分類することとした．ここで，販売数の中央値とは，61 日間の日別販売数の中央値である．

「−1」: 販売数の中央値の 0.7 倍未満の販売数
「 0 」: 販売数の中央値の 0.7 以上 1.3 倍未満の販売数
「+1」: 販売数の中央値の 1.3 倍以上の販売数

分類器には，Random Forest[2] を用いる．

3　実験

実店舗における販売履歴データを用いて本手法の有用性を示す．提案モデルでは，Random Forest の説明変数として LDA により次元圧縮した商品ごとのトピック分布と天候，曜日の One Hot ベクトルを用い，目的変数として 2.3 節で定義した $\{-1, 0, +1\}$ の 3 値を用いる．比較手法として，説明変数に販売履歴データと天候，曜日を One Hot ベクトルで表現して用いる実験を行う．

3.1　実験条件

分析対象とする販売履歴データは，中部地方を中心に展開する某小売チェーン 1 店舗において販売される商品のうち，生鮮カテゴリに属する商品とした．本研究の目的である在庫過多や品切れの解決をする際，1 日数点しか販売されない商品については対応コストの観点から優先度が高くない．そこ

で 1 日平均 10 点以上販売されたアイテムを対象とした．また，生鮮食品の多くは，1 つの材料から店舗によっては，その場で加工し提供する商品を変化させられるため，日々の需要変動に柔軟に対応可能である．

分析データの対象期間は，2014 年 10 月 1 日から 11 月 30 日の 61 日間である．分析対象期間を限定することにより，季節変動にとらわれず天候や曜日による日々の需要変動の特徴が抽出できる．商品数は 6,948，このうち 1 日平均 10 点以上販売されている生鮮食品数 I=75 を対象とする．これらの商品の総売上点数は 4,575 件である．

3.2 学習・分類手法

提案手法では LDA におけるトピック数は結果の解釈性も加味し K=10 と設定した．目的変数は需要の増減を示す $\{-1, 0, +1\}$ の 3 値とするが，これらの間に偏りが存在するためオーバーサンプリング [3] を行う．Random Forest のパラメータは，木の数を 30，木の最大深さを 20，葉の最小サンプル数を 2 とした．

検証にあたり，61 日間のデータのうち，60 日分を学習に用い，1 日分を評価に用いる leave-one-out 交差検証法を用いる．

3.3 評価方法

本研究の目的は，需要の増減を検出することによる在庫の適正化であるため，それを総合的に評価するものとして，真のカテゴリが「−1」および「+1」の商品集合に対する適合率，再現率，F 値を比較する．

3.4 実験結果

全 61 回の実験における「−1」および「+1」に対する適合率，再現率，F 値の平均の値を表 1 に示す．

表 1: 比較手法と提案手法での比較結果

	比較手法	提案手法
適合率	0.471	0.471
再現率	0.534	0.542
F 値	0.500	0.504

表 1 より，適合率では提案モデルと比較モデルは同じ値をであり，再現率と F 値において提案手法が高い値を示している．今回の実験における再現率は，対象日の需要が中央値から 30 %以上変化する商品をどれだけ正しく検出できたかを示す値である．再現率を改善できたことにより，需要変動を捉え，在庫過多や品切れに伴う機会損失の解決策がより多くの商品に対して実施可能となることから，提案手法は有効であると考えられる．

表 2 に天候ごとのトピック分布を示す．これは各日のトピック分布に対して，天候ごとに集約して平均をとったものである．また，表 3 に各商品のトピック所属確率 $P(z|x)$ 上位 5 つの商品名を示す．

表 2，3 より晴の日に所属確率の高いトピック 0 と 1 は，それぞれ所属確率 1 位が「握り寿司」である．ここで，表 2 と同様に曜日ごとのトピック分布をみると，このトピック 0，1 は平日の所属確率が高い．よって晴の日や平日には，手軽にすぐ食べられる商品が上位を占めていると考えられる．一方で，トピック 2，3 は休日および天候が曇や雨の日に高い値を示しており，野菜などの材料となる食材が多く含まれている．これらは家で調理時間が多くとれるためだと考えられる．

4 考察

本研究では LDA を適用し次元圧縮することにより，スパース性を除去することができた．そのため，次元圧縮を行わない比較モデルに対し，分類性能が向上したと考えられる．

評価の対象とした「−1」「+1」のカテゴリは全体の 4 割以上を占める．本研究で行った需要分類により機会損失を防ぐことができることから本研究の必要性がわかる．

また，天候や曜日からトピックごとの解釈を行った．この解釈と分類器の学習結果を組合せることにより，需要変動に基づくトピックも考察が可能となる．一方で，天候には気温や風量などの情報もあり，これらも活用することで，より精度の高い分類器の構築も可能となると考えられる．

5 まとめと今後の課題

本研究では販売履歴データに LDA を適用することで，各商品の需要変動パターンに対応したトピック分布を抽出し，これを用いて大幅な需要変動を検出するモデルを構築した．また，小売店での実販売履歴データを用い，需要変動の分類精度を評価し，本手法の有用性を示した．加えて，天候と曜日によって変わる各トピックの特徴に解釈を与えた．今回は，一店舗の販売履歴データを用いて評価したが，店舗による分類結果の比較や他店舗の販売履歴データを合わせて利用することによる分類精度の向上が今後の課題である．

参考文献

[1] Blei D. M., Ng A. Y., and Jordan M. I. “Latent Dirichlet Allocation,” *Journal of machine Learning research*, 3(Jan), pp.993–1022, 2003.

[2] Breiman L. “Random Forests,” *Machine learning*, 45(1), pp.5–32, 2001.

[3] Chawla N. V., Bowyer K., Hall L., and Kegelmeyer W. P. “SMOTE: Synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, 16, 2(Sep), pp.321–357, 2002.

表 2: 天候ごとのトピック分布

	トピック 0	トピック 1	トピック 2	トピック 3	トピック 4	トピック 5	トピック 6	トピック 7	トピック 8	トピック 9
晴	**0.216**	**0.197**	0.150	0.118	0.093	0.090	0.047	0.039	0.032	0.019
曇	**0.241**	0.059	0.112	**0.253**	0.118	0.111	0.030	0.022	0.045	0.008
雨	0.129	0.110	**0.183**	**0.264**	0.077	0.067	0.086	0.045	0.019	0.019

表 3: トピックごとの商品分布 $P(z|g)$

	トピック 0	トピック 1	トピック 2	トピック 3	トピック 4	トピック 5	トピック 6	トピック 7	トピック 8	トピック 9
1	**握り寿司 (細巻)**	**握り寿司 (中巻)**	生しいたけ	ほうれん草	小松菜	豚肉 (かつ)	きゅうり	ぶなしめじ	ぶなしめじ	ほうれん草
2	えのき茸	ベビーリーフ	小松菜	ニラ	馬鈴薯	ほうれん草	握り寿司 (細巻)	真だら切り身	えのき茸	ブロッコリー
3	赤ウィンナー	水菜	水菜	人参	ミニトマト	ほうれん草	豚肉 (バラ)	ちらし寿司	ウィンナー	みかん
4	大葉	寿司	小松菜	豚肉 (バラ)	コールスロー	エリンギ	ほうれん草	ほうれん草	ブロッコリー	春菊
5	ぶなしめじ	むきえび	ほうれん草	ロースハム	梨	鮭	小松菜	ごぼう	牛肉 (小間切れ)	ニラ