販売履歴データを学習した Robust Variational Autoencoder の潜在表現による店舗分析

1X17C035-1　大久保亮吾
指導教員　後藤正幸

1. 研究背景・目的

近年，フードロスの問題は，企業にとって利益追求のみならず，社会的責任という点でも重要視されている．そのため，スーパーマーケットなどの小売店においても，消費期限の短い惣菜商品の品揃え（以下，出品傾向）を検討することでフードロス解消を目指す動きが強まっている．本研究で対象とする複数店舗を有する小売チェーンにおいては，店舗規模や地域性などの店舗特性に加え，各店舗のマネージャーによる独自の意思決定により出品傾向が決定される．そのため，出品傾向の類似度が高い店舗群や独自性の強い店舗などがあると考えられ，これらを把握することで，チェーン全体でのマネジメントの一助になると考えられる．以上を踏まえ，本研究では，各店舗の出品傾向の特徴を分析することを目的とする．

本研究では，各店舗における各惣菜の出品有無を対象データの出品傾向とする．しかし，本研究で対象とする小売チェーンの各店舗で販売される惣菜の種類は，全店舗において販売されている惣菜種類数の 1 割程度に留まる．このスパース性に対応しつつ各店舗の出品傾向の特徴を比較するため，入力データの次元圧縮を行う．しかし，独自性の強い出品傾向を持つ店舗により，入力データにばらつきが大きく単純な次元圧縮手法の適用は難しい．

そこで，本研究では入力データのばらつきに頑健な次元圧縮手法である Robust Variational Autoencoder [1]（以下，RVAE）を適用する．RVAE は，深層生成モデルの一種であり，入力と出力の再構成誤差が小さくなるように学習が行われる．これにより，入力データの特徴を有した潜在表現が獲得可能となる．ここで，この RVAE の潜在表現は確率分布として出力され，一般的には，この確率分布からサンプリングされる S 個のベクトルを用いて類似度を測る．このとき，分布間の距離を正確に表すには S を十分大きくとる必要があるが，一般的には $S = 1$ としている [1]．

そこで本研究では，サンプリングすることなく確率分布間の距離を直接計算する手法を提案する．これにより，出品傾向が類似する店舗の検出が可能になる．加えて，RVAE で得られた再構成誤差から出品傾向が他の店舗と大きく異なる店舗（以下，独自店舗）の検出が可能となる．最後に，提案手法を実データに適用し，その有効性を検証する．

2. Robust Variational Autoencoder

2.1. モデル概要

RVAE は，ディープニューラルネットワークを用いて潜在表現の推論を行う深層生成モデルであり，エンコーダとデコーダから構成されている (図 1)．エンコーダでは，I 個のデータからなる入力パターン $\boldsymbol{x}_i$ $(i \in \{1, \cdots, I\})$ から確率分布を仮定した潜在表現 $\boldsymbol{z} = (z_1, \ldots, z_K)$ を推論し，デコーダでは，潜在表現から出力パターン $\tilde{\boldsymbol{x}}_i$ を生成する．エンコーダを通して潜在表現を学習し，出力パターン $\tilde{\boldsymbol{x}}_i$ が入力パターン $\boldsymbol{x}_i$ に近づくように学習を行う．

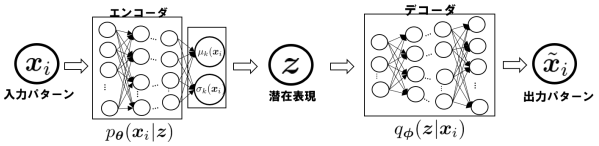

図 1: RVAE の概要図

2.2. 定式化

いま，$\mu_k(\boldsymbol{x}_i)$,$\sigma_k(\boldsymbol{x}_i)$ をそれぞれエンコーダから出力された k 番目の潜在表現の平均と標準偏差，エンコーダにおけるパラメータを $\boldsymbol{\phi}$，デコーダにおけるパラメータを $\boldsymbol{\theta}$，$p_{\boldsymbol{\theta}}(\boldsymbol{z})$ を $\boldsymbol{z}$ の事前分布，$p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z})$ を $\boldsymbol{z}$ の事後分布を近似する分布，$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)$ は入力データ $\boldsymbol{x}_i$ における $\boldsymbol{z}$ の分布とする．$\hat{p}(\boldsymbol{x}_i)$ は $p(\boldsymbol{x}_i)$ の経験分布とし D_β と D_{KL} はそれぞれ 2 つの確率分布の類似度を表す尺度である β-divergence, Kullback–Leibler divergence（KL 距離）を表すものとする．RVAE の損失関数 L は式 (1) で表され，この損失関数が最小となるよう学習を行う．

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{x}_i) = E_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)}[D_\beta(\hat{p}(\boldsymbol{x}_i)||p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}))] + D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)||p_{\boldsymbol{\theta}}(\boldsymbol{z})) \quad (1)$$

式 (1) において，第 1 項は再構成誤差と呼ばれ，第 2 項は KL 距離を表す項となっている．さらに，再構成誤差項に用いられる β-divergence により，外れ値へ頑健なモデルとなっている．

3. 提案手法

3.1. 概要

本研究では，小売チェーンにおける販売履歴データを用いて，各店舗が行う出品傾向の特徴を分析する．店舗が出品する取扱惣菜には，複数店舗に共通したものや店舗独自のものがある．そこで，店舗間の類似度を，RVAE の潜在表現により得られる確率分布間の距離で測ることを考える．確率分布間の距離は，一般的に KL 距離により算出される．しかし，KL 距離は非対称な擬距離である．そこで，本研究では，KL 距離の平均により対称性をもたせた Jensen–Shannon divergence（JS 距離）[2] を用いた算出法を提案する．

3.2. 提案モデル

各入力データ $\boldsymbol{x}_i$ に対して，RVAE より潜在表現は $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)$ の確率分布として得られる．それを用いて，データ $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ 間の JS 距離は，式 (2) から求めることができる．

$$D_{JS}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{D(\boldsymbol{x}_i||\boldsymbol{x}_j) + D(\boldsymbol{x}_j||\boldsymbol{x}_i)}{2} \quad (2)$$

このとき，$p(\boldsymbol{z}) \sim \mathcal{N}(0, I_K)$ より，$D(\boldsymbol{x}_i||\boldsymbol{x}_j)$ は以下の式 (3) から算出される．

$$D(\boldsymbol{x}_i||\boldsymbol{x}_j) = D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_j)) = \sum_{k=1}^{K} D_{KL}(q_{\boldsymbol{\phi}}(z_k|\boldsymbol{x}_i)||q_{\boldsymbol{\phi}}(z_k|\boldsymbol{x}_j)) \quad (3)$$

D_{JS} が小さいほど，各店舗の出品傾向に類似性があると考え，これにより，各店舗間の特性を分析することができる．

4. 実データを用いた分析

提案手法の有効性を示すため，提案手法を実データに適用し，分析を行う．

4.1. データ概要

分析データは，株式会社バローから提供されたスーパーマーケットの販売履歴データである．対象期間は 2016 年 1 月 1 日〜2019 年 12 月 31 日，対象店舗数は 218 店舗，対象惣菜数は 1,611 種類とした．

4.2. 分析条件

各店舗の惣菜の日別販売行列 (365 日× 1,611 種類) は，惣菜の出品がある日を 1，ない日を 0 とした．これらを 4 年 × 218 店舗分用意する．入力データは，各店舗が行う惣菜の出品傾向を定量的に表している．また，β=0.0001 とした．

4.3. 分析結果と考察

入力データに対して，RVAE を学習する．これにより，店舗ごとに再構成誤差が算出でき，これが大きな値である店舗は，独自性が高い店舗と考えられる．また，RVAE の潜在表現からデータ間の JS 距離を算出した．この距離が小さい店舗同士は類似した出品傾向を持つと考えられる．提案手法の適用により，得られた独自店舗および出品傾向が類似している店舗を示し，それらに対する考察を行う．

(1) 再構成誤差による独自店舗検出

表 1 に再構成誤差が 2σ 以上となる店舗を年別に示す．店舗名は再構成誤差の降順に並べた．これらの店舗は，他の店舗に比べて出品傾向に強い独自性を持っている考えられる．敦賀，北の森店は 4 年間通じて独自店舗として検出された．

表 1: 年ごとの独自店舗

年	店舗名
2016	**敦賀**, **北の森**, 小浜, 中曽根, 大口, 金沢元町, 本巣文殊
2017	**北の森**, 大口, 中曽根, **敦賀**, 小浜, 金沢元町, 高岡木津, 広見, 金津
2018	**北の森**, **敦賀**, 中曽根, 各務原中央, 木崎, 金津, 高岡木津, 本巣文殊
2019	**敦賀**, 各務原中央, **北の森**, 木崎, 高山, 金津, 中曽根, 羽島インター, 新田塚, 入善

(2)JS 距離による類似店舗検出

4 年間を通して独自店舗として検出された敦賀，北の森店に対し，JS 距離による近傍店舗をさらに検出する．図 2 に各年の近傍 5 店舗を示す．図中の近傍店舗には，類似した年ごとに色をつけている．

福井県にある敦賀店の類似店舗は 15 店が検出されていることがわかる．これに対して，富山県にある北の森店の類似

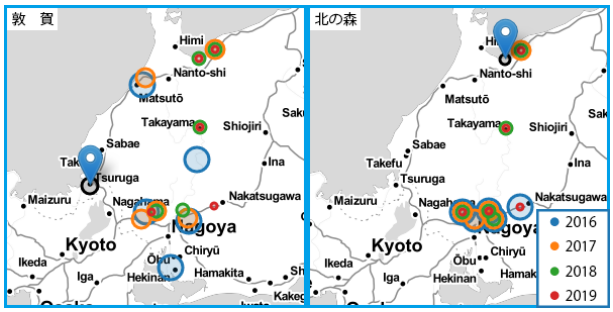

図 2: 敦賀，北の森店における類似店舗

店舗は 7 店が検出された．また，岐阜県の領下，広見店は 4 色の円となった．これは，4 年間通じて北の森店の類似店舗として検出されていることを意味する．

敦賀店が行う出品傾向は，北の森店と比べて類似店舗数が多くなることから，年ごとに独自性があると考えられる．北の森店は毎年，領下，広見店が類似店舗になるなど 4 年間で，一貫して独自性の高い出品傾向と言える．また，両店舗は，2017–2019 年の 3 年間とも類似店舗として魚津店が検出された．これより，両店舗は，魚津店を介して出品傾向に類似性があると考えられる．

(3) 出品傾向に関する詳細な考察

ある惣菜に関して，ある店舗のみが販売を行った日数（以下，惣菜別ユニーク日数）について考える．この日数が多いほど，他の店舗と異なる出品傾向であることを表す．4 年間，独自店舗となった敦賀，北の森店の惣菜別ユニーク日数では，寿司惣菜が上位に並んだ．これは，両店舗が日本海岸に立地しており，寿司に力を入れている店舗であることが要因として考えられる．一方，敦賀，北の森店の類似店舗として検出された魚津店では，寿司以外にも様々な種類の惣菜が含まれていた．3 店舗共に，日本海側に位置している．しかし，立地条件をみると魚津店は，近隣に寿司屋が多く，寿司を含んだ多様性のある出品傾向と考えられる．これに対し，敦賀，北の森店近隣には，地域に寿司屋が少なく，寿司惣菜が主力の惣菜となっていると考えられる．

以上のように提案手法を用いることで，出品傾向に基づく独自店舗，近傍店舗のみならず各惣菜の取扱商品について，詳細な分析が可能となった．

5. まとめと今後の展望

本研究では，小売チェーンの販売履歴データに対して RVAE を適用し，得られた潜在表現の確率分布間の距離により店舗特性を分析する手法を提案した．提案手法を実データに適用し，その結果について分析することで，提案手法の有効性を示した．今後の課題として，敷地面積や店舗所在地，などの店舗特有の条件を考慮したモデルの改良などが挙げられる．

参考文献

[1] H.Akrami, *et al.*,“Robust variational autoencoder,” *arXiv preprint arXiv:1905.09961*, 2019.

[2] B. Fuglede and F. Topsoe, “Jensen-Shannon Divergence and Hilbert Space Embedding,” *IEEE Int. Sym. Information Theory*, p.30, 2004.