高感度層から見る消費行動

－ディープラーニングモデルを用いて－

東海大学 情報通信学部 経営システム工学科
朝日研究室　松山 芳生

目次

1. 研究背景
2. 使用データ
3. 提案手法
    1. ディープラーニングモデルの構築
    2. 基礎集計やデータ整形, 決定木分析
4. 提案概要
    1. ディープラーニングモデルの構築
    2. 決定木分析
5. 考察
6. 今後の課題

参考文献

Appendix

1. 研究背景

- 1962年にE.M. Rogersが**イノベーター理論**[1]を提唱し、コミュニケーション・農業・公衆衛生・刑事司法・マーケティングなど、多岐に渡って成功している。

  ➡ しかし、マーケティング分野（特に日本）で<u>イノベーター理論を用いた研究</u>はあまりされていない。

- “個性的な野菜新品種導入における企業の適正”[2]

  ➡ この研究は、既存品種と異なる性質の新品種を普及させていく主体が成功しやすい企業かどうかを調べるものだった。

- 先行研究では比較的商品の販売期間が長いものを仮定している。

本研究では、比較的商品の販売期間の短いものも開発できる「食べ物」について検証する。

1. 研究背景

- イノベーター理論[1]から、イノベーター層とアーリーアダプター層に普及することができれば、アーリーマジョリティ層・レイトマジョリティ層へと繋がり、マーケットシェアが格段に上昇することが知られている。

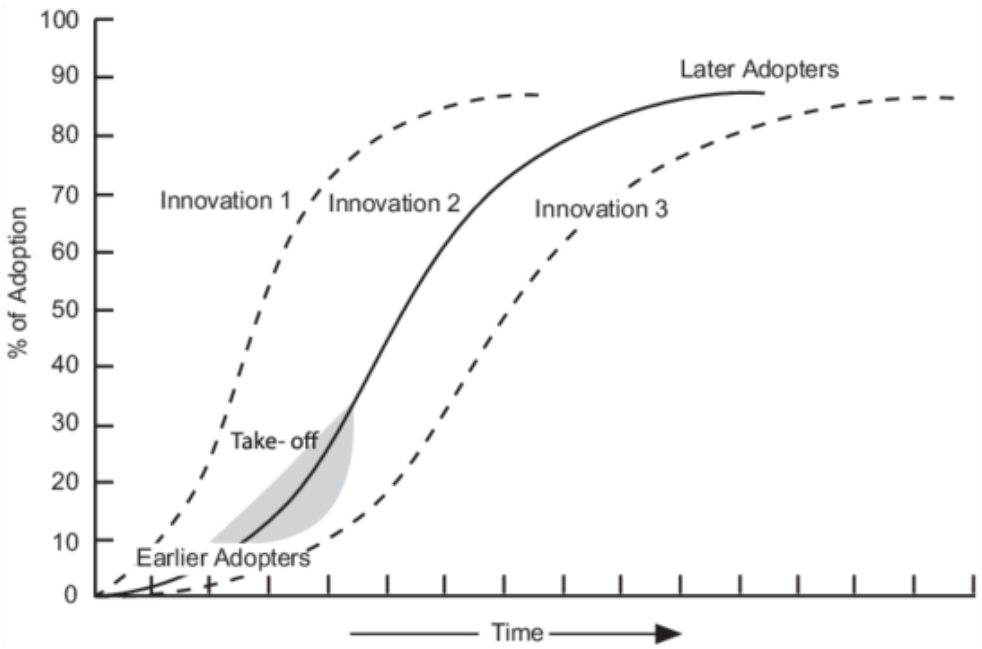


図1. 普及プロセス

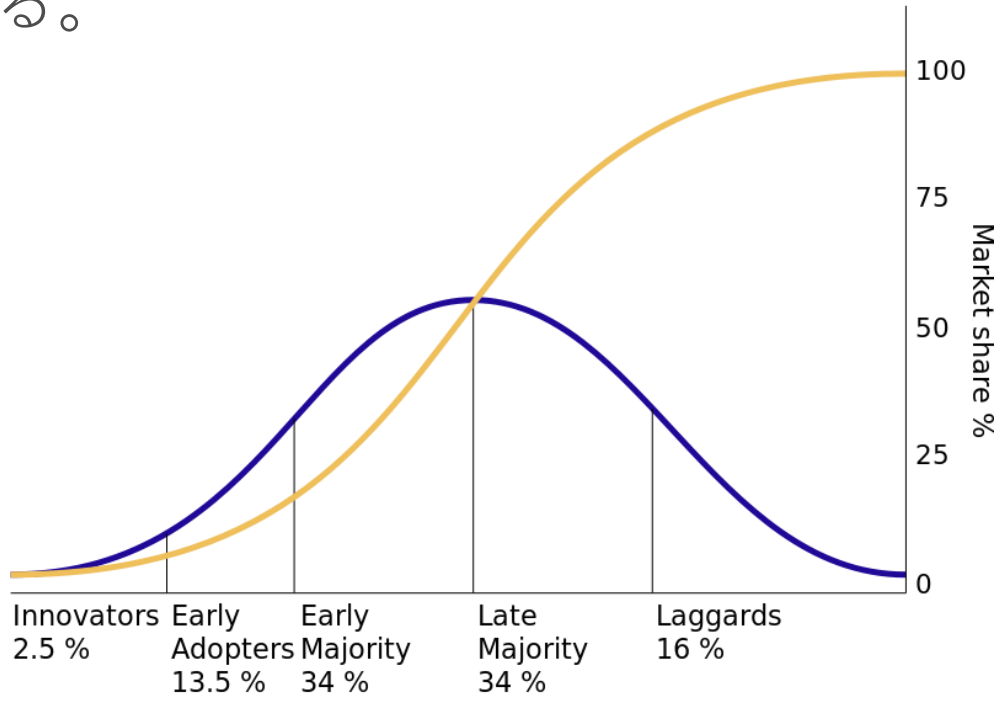


図2. カテゴリーの分類

1. 研究背景

- 本研究では、イノベーター層とアーリーアダプター層を、商品の先進性に興味を惹かれる点とトレンドに敏感で自ら常に新しい情報を収集し、他の消費者層への影響力が大きい点から**高感度層**と呼称する。

  ➡ 高いアンテナを張っている**高感度層が発信し、共感する商品**はイノベーター理論からアーリー・レイトマジョリティ層へ波及し、マーケットシェアが拡大していくと考えられる。

高感度層に多くの共感を集めるものは、イノベーター層とアーリーアダプター層に普及していると考えて、多くの共感を集める高感度消費者の特徴を掴み、この高感度消費者に対して食べ物の新商品の反応を見ることにより、**テストマーケティングの目的を果たすことができるのではないか**、と考える。

2. 使用データ

対象
生活投稿＆企業共創SNSアプリケーション「**みんレポ**」に投稿されたレポート。
今回は特にジャンルが「食べレポ」、
タイプが「買った/もらった」のレポート6,367件

収集データ期間　2016年1月1日～2016年6月30日

※ 本研究では、国立情報学研究所のIDRデータセット提供サービスにより株式会社インテージから提供を受けた「みんレポデータ」を利用した。

2. 使用データ

- みんレポ公式サイトから、“高感度な生活者から写真とコメントで商品をより魅力的に表現した、質の高い口コミ投稿が集まります。”[3]，“リアルな生活実態データ”[4]とある。

➡ **高感度層消費者のリアルな生活実態データが集まるSNS**である。

- そんな高感度消費者のみんレポユーザーの投稿「レポート」には、同じく**高感度消費者の**みんレポユーザーからの**反応**がつく。

投稿されたレポートへの反応

- **いいね**・・・・・・・レポートへの共感を示す
- ウィッシュ・・・・・レポートのブックマーク
- アクティビティ・・・自分も買った、食べたなどを示す

3. 提案手法

- **ディープラーニングモデルを構築する。**

  ➡ 教師あり学習でディープラーニングモデルを構築することによって汎用性の高いモデルを作成する。

- **ディープラーニングモデルを適用し、教師なしデータを予測する。**

  ➡ 予測されたデータは、標本の特徴に依ることなく予測され、より高感度層の中の共感されやすいレポートを書く人の特徴が浮き彫りになる。

  ➡ 予測されたデータについて**決定木分析**をかけることにより、特徴を視覚的に浮き彫りにする。

3.1 **提案手法** ディープラーニングモデルの構築

- ディープラーニングモデルの構築にはDeep Learnerを使用した。

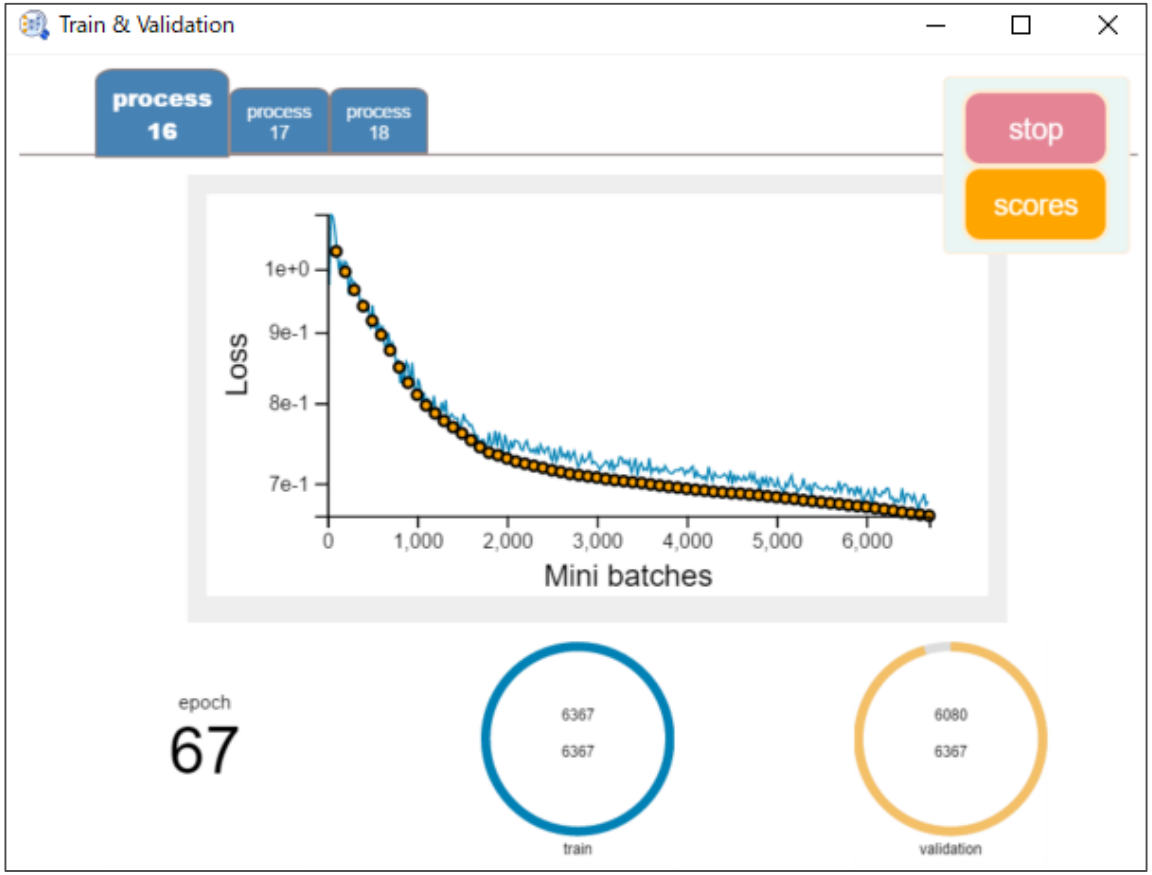


図3. DeepLearnerでの学習の様子

3.1 **提案手法** ディープラーニングモデルの構築

- Deep Learnerではプログラミングの知識が無くても、
  アイコン操作でディープラーニングを実行することができる。

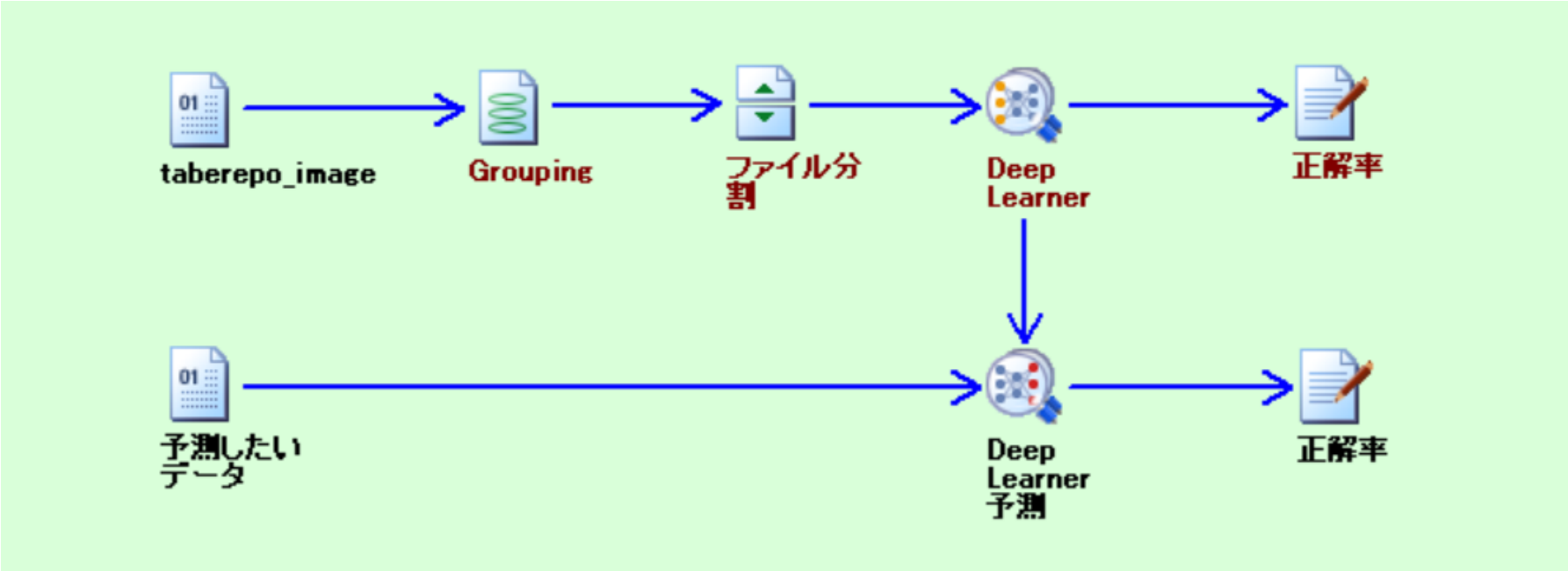


図4. ディープラーニングの学習フロー図

- 学習させたモデルから、**未知のデータの予測**も可能。

3.2 **提案手法** 基礎集計やデータ整形, 決定木分析

- 標本集団の分布の確認・基礎集計・定量値の2値カテゴリ化
  ・モデル構築用データの整形, 決定木分析をするために
  Visual Mining Studio(VMS)を使用した。
  VMSは簡単な操作で本格的なデータマイニング、分析前のデータ加工を行えるツールである。

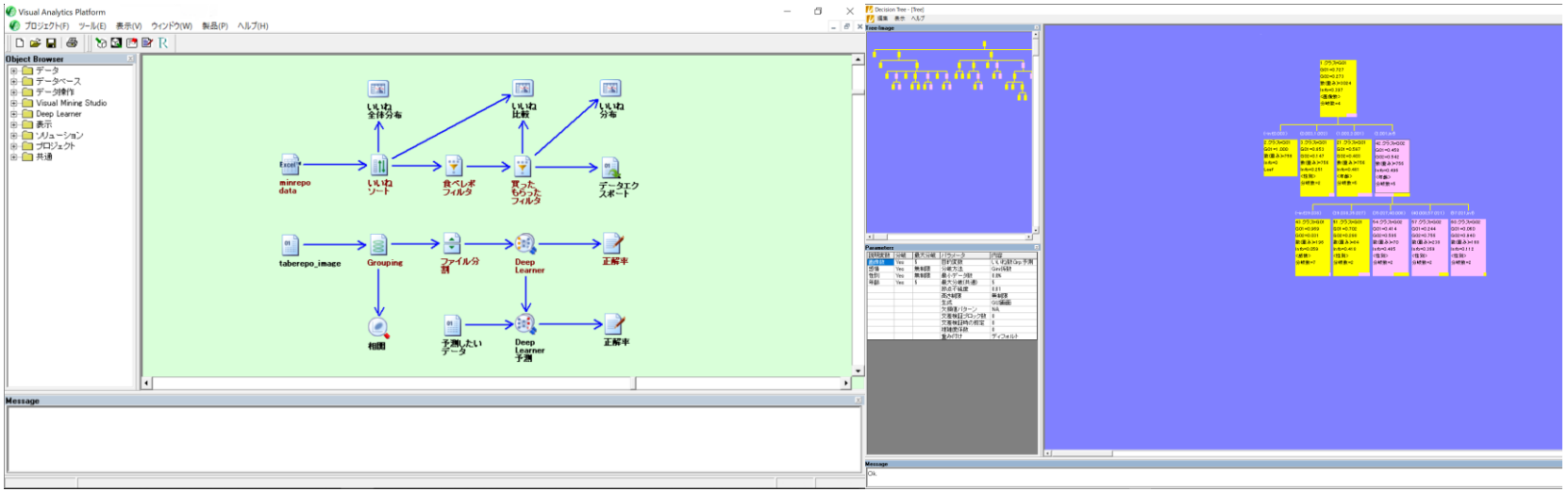


図5. VMSの利用例

4.1 **提案概要** ディープラーニングモデルの構築

- 投稿されたレポートに対しての共感を表す、「**いいね**」を**目的変数**にとる。

みんレポアプリの人気タブでは、3日以内に投稿されたレポートで「いいね」が多い順にソートされて紹介される。
1度に読み込まれるレポートは、「いいね」が10以上のものが占めている。

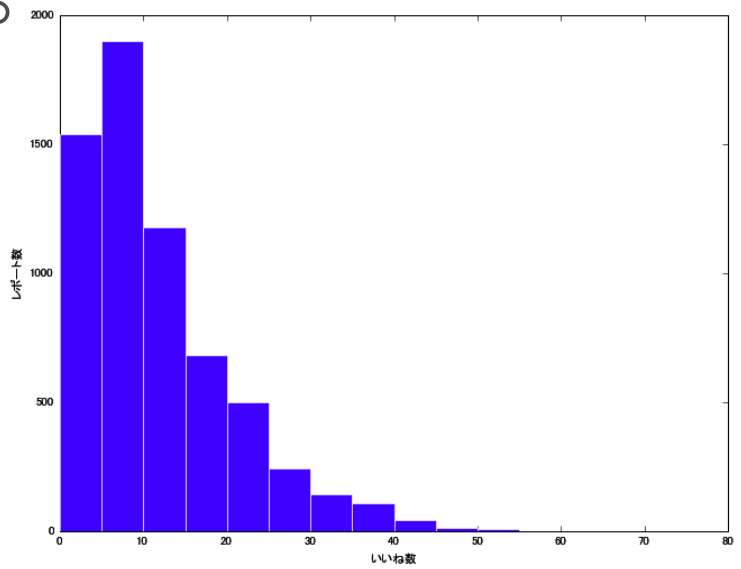


図6. レポート毎のいいね分布

「いいね」が10未満のレポートを「**非共感レポ**」、
「いいね」が10以上のレポートを「**共感レポ**」とする。

4.1 **提案概要** ディープラーニングモデルの構築

- 本研究で作成するディープラーニングモデルは、
  教師あり学習で予測対象が2値カテゴリの**分類モデル**である。

- 扱うデータから、ディープラーニングモデルの構築に
  多層パーセプトロンというネットワーク構造を用いる。

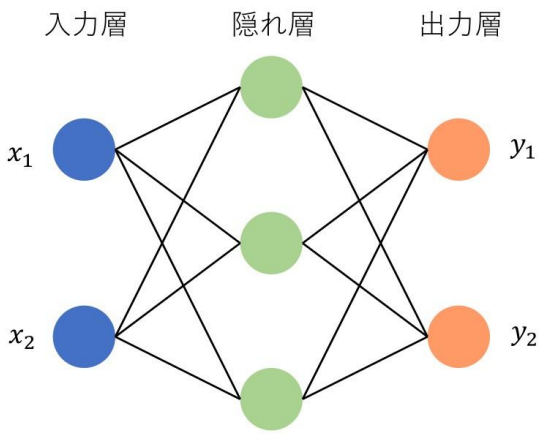


図7. 多層パーセプトロン
(隠れ層：1)の構造

図8. 今回のモデル構造

- 本研究のモデルは、
  隠れ層が2つの構造と
  なっている。

4.1 **提案概要** ディープラーニングモデルの構築

● 予測対象（**目的変数**）は「非共感・共感レポ」の2値カテゴリ。

説明変数は

- 添付された**画像数**
- レポを投稿したときの**感情**
- レポを投稿した人の**性別**
- レポを投稿した人の**年齢**

の**4つ**を選択。

レポートへの反応として

- ウィッシュ
- アクティビティ

の**2つ**は「いいね」されて
共感したレポに対して
付くものであり、
これらは説明変数に
入れることはしない。

4.1 **提案概要** ディープラーニングモデルの構築

- 説明変数について

  ディープラーニングモデルの構築において、
  - モデルの複雑さの回避
  - 過学習の防止

  をするために説明変数は汎用性を失わないように選択した。

- 多重共線性について

  吉田(1987)[6]によると、多重共線性は説明変数間に相関がある場合に相関行列が正則でなくなることで解が求められなかったり、不安定であったりすることが本質的な問題であると言われている。

➡ Deep Learnerでは、Dropoutによる正則化を行っている。
これにより過学習の防止[7]だけでなく、
多重共線性についても防止されている。

4.1 **提案概要** ディープラーニングモデルの構築

- Deep Learnerでの学習には、全6,367件のレポートのうち、ランダムで8割(5,094件)を学習用に、2割(1,237件)を検証用に使用した。

- 隠れ層の数は任意で指定することができるが、浅川(2014)[8]によると、少なくとも2つの隠れ層が必要であると言われている。
  また、Nielsen(2014)[9]によると、隠れ層の数が多くなるほど確率的勾配降下法では勾配消失問題・勾配の不安定化が生じると言われているので、隠れ層の数を増やしすぎてもいけない。

➡ 以上より、隠れ層の数は**2つ**にした。

4.1 **提案概要** ディープラーニングモデルの構築

- 多層パーセプトロンの各隠れ層では、
  - 出力次元数
  - 活性化関数
  - Dropout Ratio

  の3つのハイパーパラメータを設定できる。

- Deep Learnerでは、ディープラーニングを複雑化しているパラメータ調節をModel Optimizer機能により**自動**で調節してくれる。

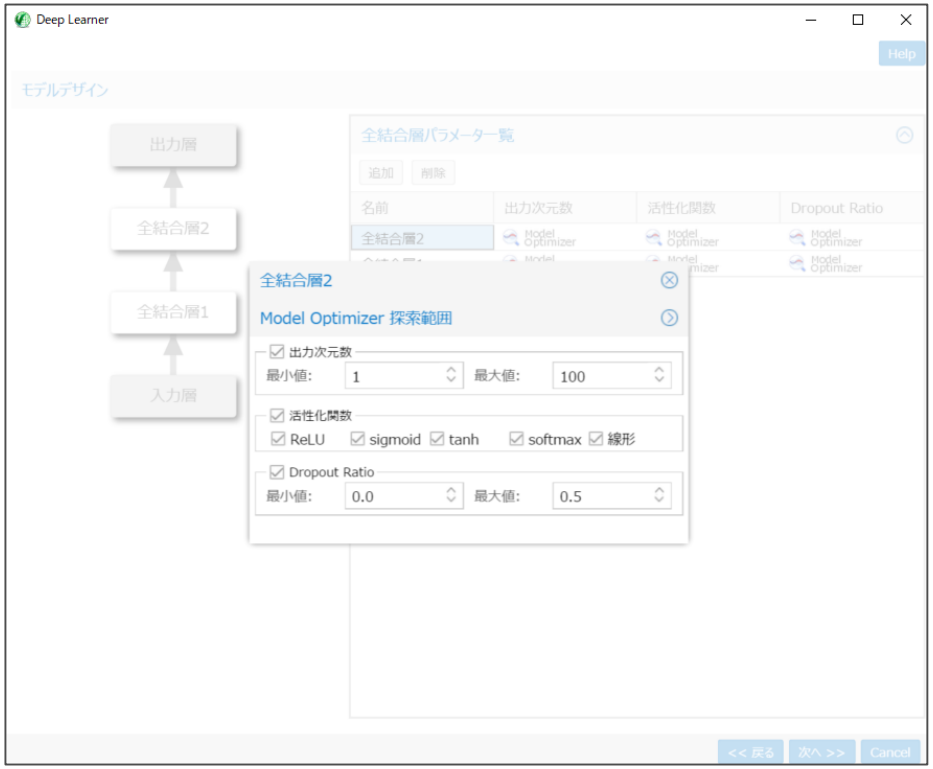


図9. 隠れ層のパラメータ調節

4.1 **提案概要** ディープラーニングモデルの構築

- Model Optimizerの設定

  Model Optimizerでは、自動でパラメータを調節してくれるが、
  その調節する方法は自分で設定しなくてはならない。

  - **手法**についてはランダムとベイズ最適化の
    2つがあるが、本研究ではRandom Forestを利用した
    ベイズ最適化を選択した。

  - **指標**についてはLoss, Accuracy, F1-score, Precision, Recallの
    5つあるが、本研究で判別したい2値カテゴリは予測結果が
    正解か否かで評価していいものなので、**Accuracy**を選択した。

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定

  Model Optimizerでは調節されない最適化関数・学習率・エポック数・ミニバッチサイズについても自分で調節する必要がある。

  - **最適化関数**については、Deep LearnerにはAdam, SimpleSGD, AdaDelta, AdaGrad, RMSpropの5つが用意されている。
    しかし、データによって最適な最適化関数は変わるため、
    試してみるまで分からない。ここで、5つの最適化関数以外の
    学習設定を同じにし、正解率の高い最適化関数を選択する。

    ➡ 下記表1より、本研究においては**AdaDelta**を選択。

表1. 最適化関数ごとの正解率

最適化関数	Adam	SimpleSGD	**AdaDelta**	AdaGrad	RMSprop
正解率	**0.72536**	**0.68**	**0.73263**	**0.64**	**0.72**

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定

  - **学習率**については、田口(1998)[10]によると、大きくすると誤差の収束が速くなるが、不安定になる。小さくすると誤差の安定性は増加するが、収束までに時間がかかると言われている。
    これもデータによるので、学習率以外の条件を同じにし、
    比較・検証する。

    ➡ 下記表2より、本研究データにおいては
    学習率を大きく変えても正解率に変動がなかったため、
    デフォルト値の**0.0001**にする。

表2. 学習率ごとの正解率

学習率	0.001	0.0001	0.00001
正解率	**0.73263**	**0.73263**	**0.73263**

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定
  - **ミニバッチサイズ**は、学習データをその数からなる部分に分割し、各ミニバッチについてその値とパラメータに対する勾配を計算し、更新式(1)を w で更新するもの。（※ Deep Learner技術資料より）
    数が少ないほど分割数が多くなるので、各ミニバッチに含まれるデータの影響が大きくなる。

➡ **小さくなるほど過学習を起こしやすい。**

$$L(x;w) = \sum_{i=1}^{N_{data}} L^{(i)}\left(x^{(i)};w\right) \qquad (1)$$

$L(x;w)$ = 損失関数，x = 入力，w = パラメータ，N_{data} = 全データ

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定
  - ミニバッチサイズについても、過学習を起こさない程度に少なめの値をとることが望ましい。さらに、適度なミニバッチサイズを選ぶことができれば、並列計算による高速化とメモリの節約も達成できる。（※ Deep Learner技術資料より）通常、2^n個に分けることが多い。これにより、メモリアクセス効率が良くなるからである。

➡こちらについても**ミニバッチサイズ**は比較・検証を行い、最終的にデフォルト値(**64**)に落ち着いた。

表3. ミニバッチサイズごとの正解率

バッチサイズ	32	64	128
正解率	**0.73380**	**0.73263**	**0.72556**

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定
  - **エポック数**は、1つの学習(訓練)データを繰り返して学習させる数のことである。ディープラーニングでは訓練データを繰り返して学習することでパラメータを学習するため、過学習を起こさない程度に多くエポック数をとらなければならない。
    そんなエポック数の決め方に**Early Stopping**という方法がある。

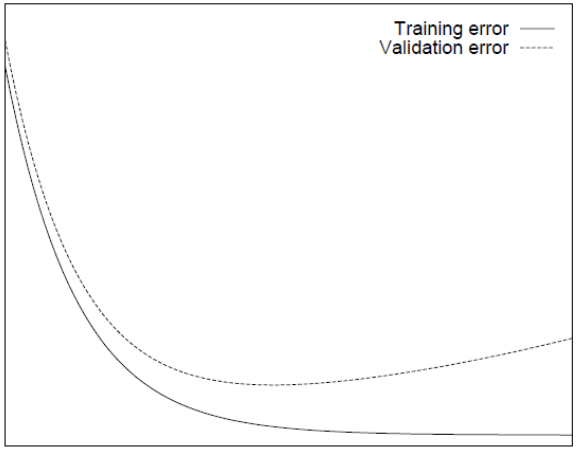


図10. 誤差の経時変化[11]
(x軸:時間, y軸:誤差)

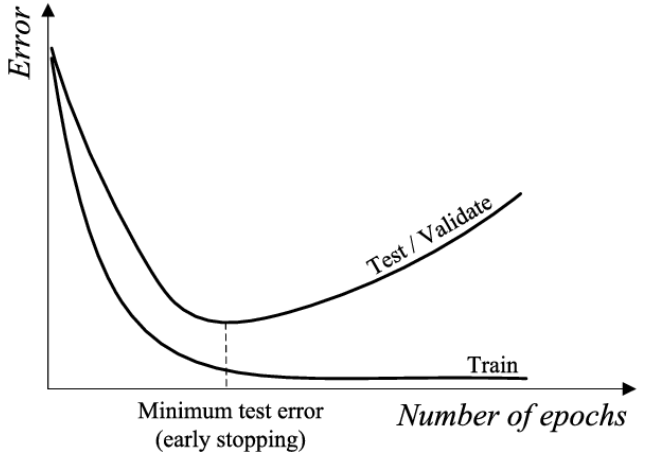


図11. 誤差とエポック数の関係[12]

4.1 **提案概要** ディープラーニングモデルの構築

- 学習設定
  - **Early Stopping**とは、誤差が収束し始めたところで学習を止めることである。これにより**過学習をしないように学習を止めることが可能**になる。ページ22の図11の”Test/Validate”の誤差が大きくなっていっているところが過学習を起こしている。
    - Deep Learnerでは、学習している様子を逐次確認できる。
      この画面では、誤差が収束してきたら学習を止められる。

      ➡ 本研究では、**100**に固定する。

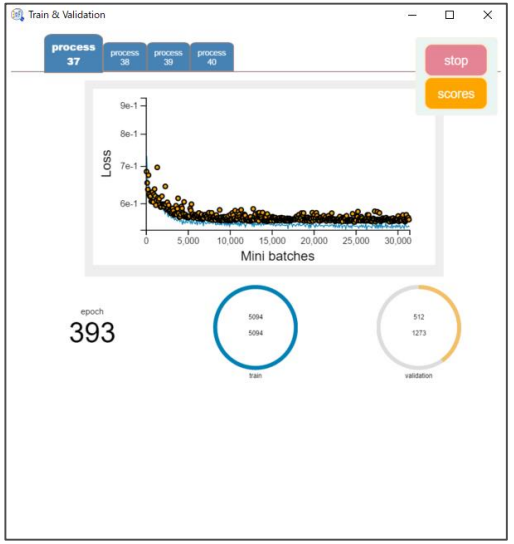


図12. 誤差が収束している様子（過学習）

4.1 **提案概要** ディープラーニングモデルの構築

● 以上の設定を行ったディープラーニングモデルを構築し、

- 添付された**画像数**
- レポを投稿したときの**感情**
- レポを投稿した人の**性別**
- レポを投稿した人の**年齢**

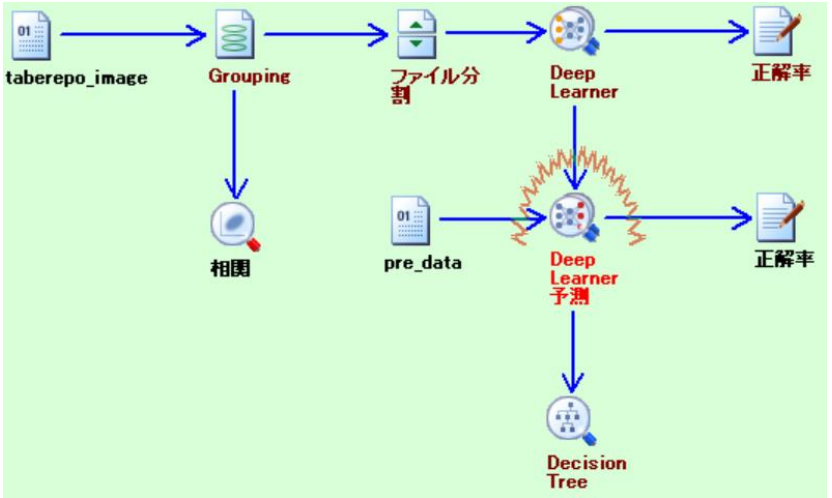


図13. 学習したモデルを適用した予測

教師なしの上記4つの変数を持ったデータが存在したとき、
構築したディープラーニングモデルを用いてレポートを書いた人が、
「非共感レポ」・「共感レポ」のどちらを書くのかを**予測**する。

4.2 **提案概要**　決定木分析

- 本研究で用いた説明変数は4種類である。

表4. 説明変数ごとに取れる値

画像数	感情	性別	年齢
0	なし	男性	16
1	いいね	女性	17
2	感動		・
3	喜び		・
	困った		・
	残念		68
	怒り		69

  - 表4より、本研究では
    説明変数が取れる値が、
    $4 \times 7 \times 2 \times 54 = 3{,}024$
    の3,024通りしかない。

➡ 全パターンの組み合わせのデータを作成し、そのデータについて学習させたディープラーニングモデルを適用し、「非共感レポ」・「共感レポ」のどちらを書くのかを予測し、決定木分析により予測された2値の特徴を分析する。

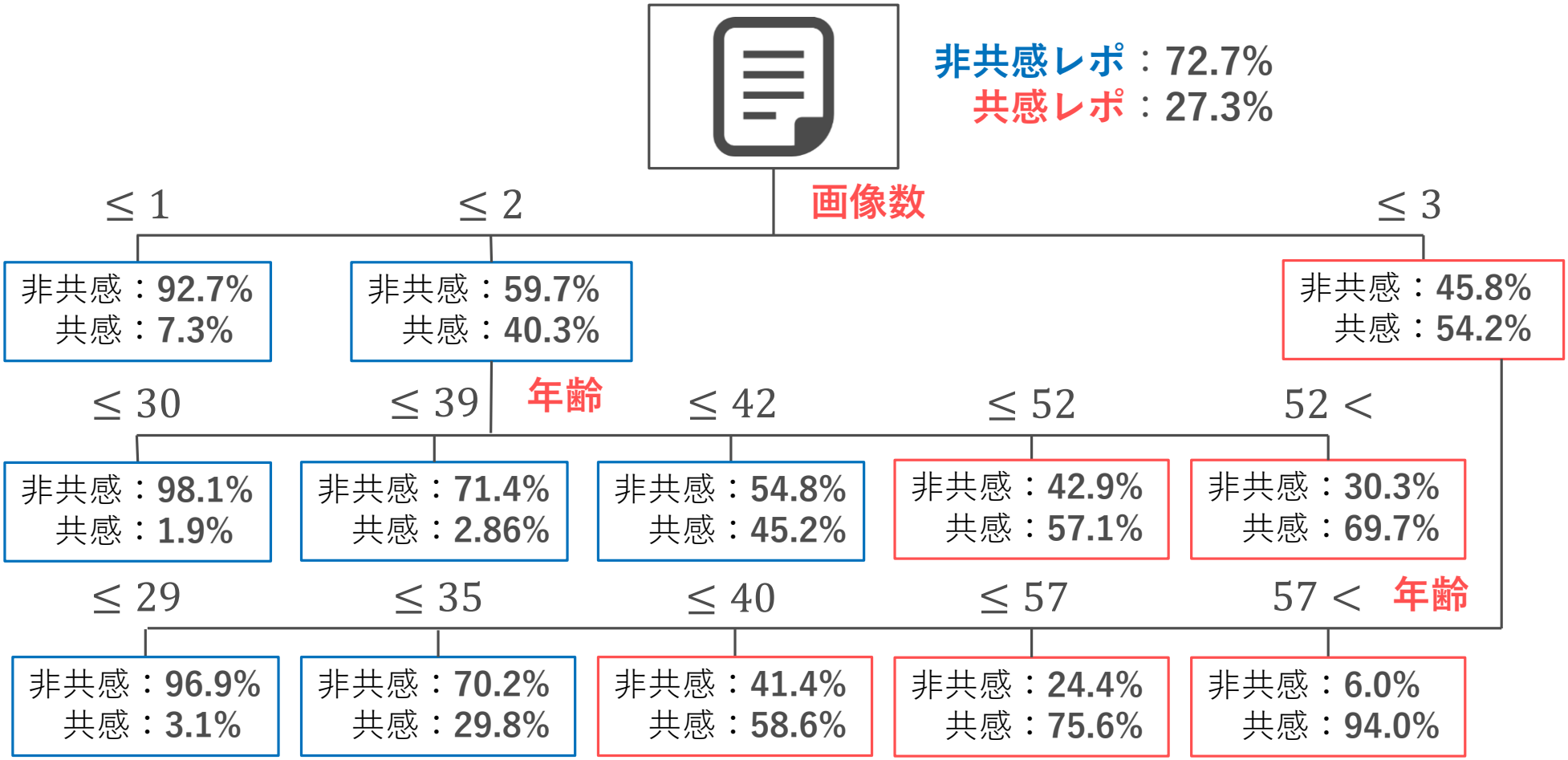
4.2 **提案概要**　決定木分析

- 分析条件

表5. Decision Treeオプション設定値

説明変数	分岐	最大分岐	パラメータ	内容
画像数	Yes	5	目的変数	いいね数.Grp.予測
感情	Yes	無制限	分岐方法	Gini係数
性別	Yes	無制限	最小データ数	0.80%
年齢	Yes	5	最大分岐(共通)	5
			節点不純度	0.01
			高さ制限	無制限
			生成	GUI画面
			欠損値パターン	NA,
			交差検証ブロック数	0
			交差検証時の剪定	0
			複雑度係数	0
			重み付け	ディフォルト

4.2 提案概要　決定木分析



5. 考察

- 決定木分析の結果より、**共感されやすいレポート**は、第一に画像が多いほど共感されやすいことが分かった。画像が0, 1枚しか添付されていないレポートは92.7%も非共感レポートに分類されていて、**共感レポートの第一条件として画像が2枚以上必要**である。やはり視覚的な情報は単なるテキストデータよりも情報量が多く、共感を得やすいことが分かる。

- 次に、画像が2枚添付されているレポートは年齢が43歳以上、画像が3枚添付されているレポートは年齢が36歳以上であることが分かる。添付画像が2枚以上ある条件のもと、**年を重ねている人ほど共感を得やすい**ことが分かる。決定木分析の結果からも年齢を重ねるほど、共感レポートの割合が増えていることが分かる。こちらの条件は感覚的にも理解しやすい。

5. 考察

- 逆に、画像数が2枚でも3枚でも30歳以下では実に97%以上が非共感レポートと分類されている。画像数が2枚だと39歳以下で、画像数が3枚だと35歳以下でも70%以上が非共感レポートと分類されている。共感されやすいレポートを書くのは、いわゆる**アラフォーと言われる世代から**と分かる。
- **説明変数**には画像数と年齢のほかに感情と性別が入れられている。本研究で作成したモデルからは、**感情と性別はそれほど重要な条件とは認められなかった**。

➡ 食べ物の新商品を作る際は、イノベーター理論から高感度層を気にする必要がある。また、本研究の結果より、高感度層の中でも画像を多く添付する、年を重ねている人について反応を見ることにより、検証は必要だが、**テストマーケティングの役割に耐えうると考える。**

6. 今後の課題

- 今回は、「非共感レポ」・「共感レポ」の2値を予測するディープラーニングモデルを構築したが、その説明変数には数値的に解釈できるもの(名義尺度を含む)に限られている。人々にどのような言葉が共感されやすいのか、テキストマイニングからディープラーニングモデルを構築すると、更に精度が上がると考えられる。

- また、今回のデータには画像ファイルも含まれていた。こちらもAIなどにより画像タグ付けを行ってから、テキストマイニングを行うことによって、どのような画像が共感されやすいのか、決定木分析の結果からも画像の重要性が示されているため、こちらについても検証したい。

参考文献

[1] E.M. Rogers: *Diffusion of innovations the third edition*, macmillan publishers (1962)
[2] 森尾昭文, “個性的な野菜新品種導入における企業の適正”, 中央農業総合研究センター研究報告誌(17), pp.39-48 (2012)
[3] 株式会社インテージ みんレポキャンペーン
https://www.intage.co.jp/solution/process/product-development/minrepocp/
(最終閲覧日：2018/10/20)
[4] 株式会社インテージ みんレポSNS分析
https://www.intage.co.jp/solution/process/market/minreposns/
(最終閲覧日：2018/10/20)
[5] Frank. Rosenblatt: *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological review, Vol.65, No.6, pp.386-408 (1958)
[6] 吉田光雄, “重回帰分析における多重共線性とRidge回帰について”, 大阪大学人間科学部紀要, Vol.13, pp.227-242 (1987)
[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov: *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research, pp.1929-1958, Vol.15 (2014)

参考文献

[8] 浅川伸一, “ディープラーニングと中間層の意味”, 日本認知心理学会第12回大会書誌, pp28, (2014)
[9] Michael A. Nielsen: *Neural Networks and Deep Learning*, Determination Press (2014)
[10] 田口功, “バックプロパケーション・ニューラルネットワークにおける収束条件と学習の加速化”, 敬愛大学国際研究第2号, pp77-109, (1998)
[11] Lutz Prechelt: *Early Stopping – but when?*, Fakultät für Informatik; Universität Karlsruhe, pp55-69 (1998)
[12] Jorge M. Santos: *Data classification with neural networks and entropic criteria*, Universidade da Beira Interior Tese de doutoramento (2007)

Appendix

説明変数間の関係

- 14ページで正則化を行うことにより多重共線性は
  防止されているとしているが、説明変数間の関係を見ておく。

- 説明変数の尺度より、
  「**順序**」と「**比率**」ではスピアマンの順位相関係数を、
  「**名義**」と「**順序**」・「**名義**」と「**比率**」では相関比を、
  「**名義**」と「**名義**」ではクラメルの連関係数を用いて
  説明変数間の関係を見る。

表6. 説明変数の形式

説明変数	形式
画像数	順序尺度
感情	名義尺度
性別	名義尺度
年齢	比率尺度

表7. 説明変数間の関係値

変数1	変数2	**順位相関係数**
画像数	**年齢**	**0.15**

変数1	変数2	**独立係数**
感情	**性別**	**0.07**

変数1	変数2	**相関比**
感情	**画像数**	**0.04**
感情	**年齢**	**0.06**
性別	**画像数**	**0.00**
性別	**年齢**	**0.13**

説明変数間の関係

- 前ページ表6, 表7より、もし正則化による多重共線性の防止が行われていなかったとしても、説明変数間の関係はほとんどないものとみなせるので、多重共線性の恐れが少ないと言える。

- 正則化による多重共線性の防止が可能ならば、
ディープラーニングモデルでは説明変数にどんなものを入れてもそれなりの精度で予測が可能であることを示している。
もちろん、欠損値や異常値については極力対処する必要がある。