学習者の心的状態推定モデルが獲得した潜在的分類観点と判断根拠抽出の試み

Extraction of Latent Classification Criteria and Inference of Multi-layered Neural Network Estimating Learner's Mental States

古澤 嘉久[1*]　田和辻 可昌[1,2]　松居 辰則[2]
Yoshihisa FURUSAWA[1]　Yoshimasa TAWATSUJI[1,2]　Tatsunori MATSUI[2]

[1] 早稲田大学 大学院人間科学研究科
[1] Graduate School of Human Sciences, Waseda University
[2] 早稲田大学 人間科学学術院
[2] Faculty of Human Sciences, Waseda University

Abstract:
We constructed deep neural network system estimating learner's mental states during a class. However, its classification criteria – that is, how the system estimated the mental states – is not interpretable. In this research, we tried to extract the criteria with two methods; factor analysis and saliency map. The result of factor analysis indicated that the system obtained three latent axis to classify the mental states. The result of saliency map indicated that not only physiological data but also teachers' utterance play an important role to estimate mental states of "Shame," "Hopeless," "Boredom," and "Relief."

1　はじめに

教授・学習過程において学習者の心的状態を把握することは，教育効果・学習効果の観点から極めて重要である．我々は，これまで教師と学習者のインタラクションにおいて，教師の発話と学習者の生体情報から学習者の心的状態を推定する学習器モデルを構築してきた [1]．ところが，この学習器がどのような観点から学習者の心的状態を分類しているのかについては十分検討がなされていない．

近年，深層学習を初めとした機械学習器が「何を学習したか」といった解釈性・説明性に関して国内外問わず広く議論が進められている [2]．このような機械学習器における解釈性については，Grad-CAM [3] などに代表されるように広く画像の分類問題で議論が進められてきた．一方で，本研究で構築された機械学習器は，入力を画像情報としない学習環境におけるデータであり，本ドメインでの学習器の解釈性・可視化は今後重要な課題となると考えられる．そこで本研究では，このような教育ドメインのデータを用いて構築された機械学習器が「何を学習したか」という点を多角的に分析し，本アプローチの有効性を検討することを目的とする．

*連絡先：早稲田大学 大学院人間科学研究科
〒 359-1192 埼玉県所沢市三ヶ島 2-579-15
E-mail: f.y_1996_w-skk@akane.waseda.jp

2　研究の観点とアプローチ

機械学習器が学習によって獲得したと考えられる分類観点を解釈する際には，どのような観点からこのような出力を表出したか，という方向の視点が重要である．心理学において，このような方向の視点（本来は観察することができない潜在的な要因が出力に影響を与えているという視点）から，人間の行動を説明する際に因子分析が広く採用されている．そこで，機械学習器全体を black box ととらえた際に，潜在的にどのような観点から出力がなされたのかという点から出力データに対して因子分析を適用することを検討した．一方で，本研究で用いた機械学習器である多層ニューラルネットワークは，各ノードとそれらを結合する重みによって数学的に表現されており，出力に対してどの入力が寄与しているかを解析的に計算することが可能である．以上二つの観点から機械学習器を分析することによって，機械学習器が「何を学習したか」について多角的に考察することが可能であると考えられる．

3　心的状態推定器

学習器の構築に用いたデータは松居ら [4] によって行われた実験において測定・観察された情報のうち，学習者の皮膚コンダクタンス，容積脈波，呼吸強度および教師の発話カテゴリデータである．ネットワークの構造は，中間層 4 層からなる多層ニューラルネットワークであり，入力層では，生体情報（皮膚コンダクタンス，容積脈波，呼吸強度）および教師の各発話カテゴリに対応したベクトルからなる．また，出力層は，学習者の心的状態カテゴリ 8 種類（Enjoy, Hope, Pride, Anxiety, Shame, Hopeless, Boredom, Relief）に対応した 8 次元のベクトルからなる．表 1 に本実験で用いたネットワークの構造を示す．

表 1: 構築したネットワークの構造

層数	6 層（中間層 4 層）
中間層のユニット数	1 層目から 32, 64, 64, 32
最適化手法	Momentum SGD
学習方法	誤差逆伝搬法
活性化関数	中間層：ReLU 関数 出力層：Softmax 関数

学習にあたってはデータを 6:4 に分割し，データの 60

図 1 に Accuracy と Loss の変化を示す．横軸は epoch 数，縦軸はそれぞれ Accuracy および Loss を示している．どちらも epoch を経るにしたがって収束していることが分かる．また，各心的状態における正答率は訓練データにおいて，Enjoy が 73.9%，Hope が 72.7%，Shame が 70.3%，Hopeless が 65.6%の精度で，検証データにおいて Enjoy が 72.2%，Hope が 70.9%，Shame が 68.4%，Hopeless が 62.3%の精度で，それぞれ正しく推定されていた．これらはいずれも訓練・検証データにおいてもデータ数が他の心的状態のデータ数よりも多く，このことから高い精度で分類することが可能であったと示唆される．

4　潜在的分類観点の抽出

潜在的な分類観点の抽出方法として，本研究では因子分析を用いる．まず因子分析の概要について述べ，その後実験データから構築した学習器の出力情報に対して因子分析を適用した結果について述べる．

4.1　因子分析

因子分析は統計的分析手法のひとつで，データの背後にある要因（共通因子と呼ばれる）を推定する手法である．実験等で得られた観察データを $\boldsymbol{t} \in \mathbb{R}^p$，共通因子を

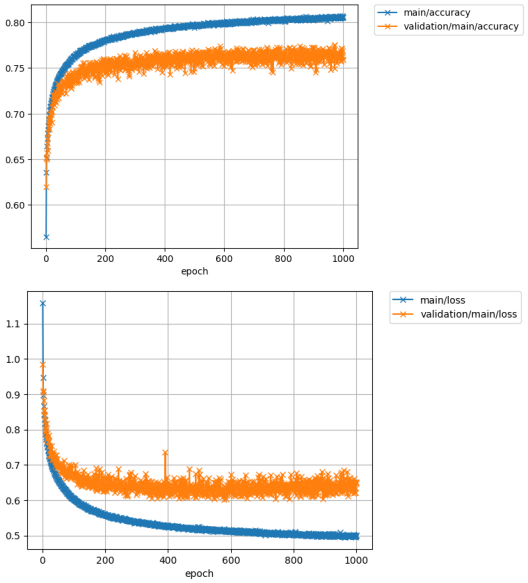

図 1: 学習器の Accuracy（上段）と Loss（下段）の変化

$\boldsymbol{f} \in \mathbb{R}^n$ とし，共通因子が観察データに与える影響の程度を要素としてもつ因子負荷行列を $A = (a_{ij}) \in \mathbb{R}^{p \times n}$ とすると，

$$\boldsymbol{t} \quad = \quad A\boldsymbol{f} + \boldsymbol{e} \tag{1}$$

の線形モデルを仮定した分析手法である．ここで，$\boldsymbol{e}\mathbb{R}^p$ は観察データの各要素にのみ影響を与える独自因子である．因子分析モデルは，主に心理学などの分野で SD 法の一環として用いられ [5]，実験参加者間に共通して存在する因子（能力や評価軸）を抽出する際に用いられる．

4.2　学習器が獲得した評価軸の抽出

因子分析の手法から，学習器が獲得した潜在的な評価観点を抽出することを試みた．今回，訓練データにおける出力層の出力値（Softmax 関数を適用する前のデータ）に対して因子分析を適用した．図 2 に平行分析を使用したスクリープロットを示す．赤色が出力データから構築された相関行列の固有値を昇順に並べたものである．また，青色が出力データと同次元で，要素が標準正規分布から無作為抽出された値で作成されたデータから構築された相関行列の固有値を昇順に並べたものである．この結果から，2 因子が検討されたが，2 因子での因子分析の結果，Boredom の独自性が極めて高かった（$u_{boredom} = 0.847$）ため，再度因子数を 3 として因子分析を行った．

この結果得られた因子負荷行列を表 2 に示す．因子 1 で説明される感情は，Enjoy, Hope, Anxiety, Shame,

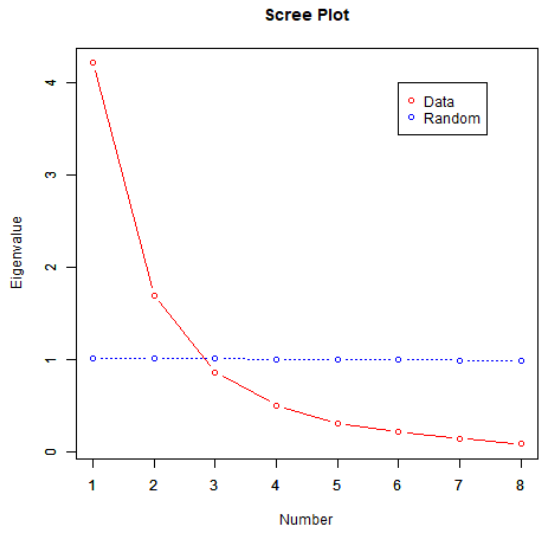

図 2: スクリープロット（赤色：本データ，青色：ランダムデータ）

Pride であり，因子 2 で説明される感情は Hopeless, Relief であった．また，因子 3 は Boredom に対して強く寄与した因子であった．

表 2: 因子負荷行列

感情	因子 1	因子 2	因子 3	共通性
Enjoy	1.04	0.13	-0.22	0.96
Hope	0.99	-0.19	-0.09	0.84
Anxiety	-0.83	0.31	-0.14	0.78
Shame	0.73	0.29	-0.05	0.67
Pride	-0.56	-0.04	-0.21	0.51
Hopeless	-0.23	0.98	-0.07	0.9
Relief	-0.23	-0.66	-0.16	0.67
Boredom	-0.11	0.03	1.04	1

次に，得られた因子負荷行列および入力データの因子得点を因子空間上にマッピングした結果について述べる．図 3 に，各因子の組ごとの 2 次元因子空間上にマッピングした因子負荷行列の要素および因子得点を示す．この結果から，各因子得点は因子空間上の原点に集中していることが分かる．また，因子 1 と因子 2 のうち，正の方向に位置づけられている Enjoy，Hope，Shame（因子 1）および Hopeless（因子 2）はそれぞれ，訓練データ，検証データともに正答率が高い心的状態であることが分かる．これらのことから，各因子に対してポジティブに寄与している心的状態はおおむね学習データに多く含まれるものであり，データの性質と因子負荷量の符号の間に関係があることが示唆される．

5　出力に寄与する入力の解析

機械学習器における可視化研究では，出力層の各ユニットに対してどの入力が寄与しているかは重要である．本節ではまず可視化に関する研究手法について述べ，本研究で行った解析について述べる．

5.1　可視化に関する関連研究

入力画像 $\boldsymbol{x}$ からクラス c を予測するモデルにおいて，Simonyan [8] は，$S_c(\boldsymbol{x})$ が線形的なモデルの場合は，入力画像に対する重要さとはそのモデルの重みとして表現されるとして，下記のように，クラス c へ連結する重み $\boldsymbol{w_c} \in \mathbb{R}^n$ が重要さに該当していると主張している．ただし，$b_c \in \mathbb{R}$ はバイアス項である．

$$S_c(\boldsymbol{x}) = \boldsymbol{w_c^T x} + b_c \tag{2}$$

また，実際のモデル $S_c(\boldsymbol{x})$ は非線形であるため，入力画像に関してテイラー展開し，一次近似した際に，以下のように入力画像での勾配が重み部分に該当することがわかる．

$$S_c(\boldsymbol{x}) \approx \boldsymbol{w^T x} + b, \quad \boldsymbol{w} = \left.\frac{\partial S_c(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x_0}} \tag{3}$$

そこで Simonyan[8] は，最終的なクラス c への出力から逆伝播を行い，入力画像に逆伝播されてきた勾配を Saliency Map として採用することを提案している．またこれは，入力画像の個々のピクセルにおける微小変化における $S_c(\boldsymbol{x})$ の変化を表しており，微小変化を及ぼした際に $S_c(\boldsymbol{x})$ が大きく変化したものほど，$S_c(\boldsymbol{x})$ にとって重要であるという見方ができる．

一方で，Erhan[9] らは，入力とネットワークの出力の関係を調査するために，ある特定のユニット (今回の場合はソフトマックスをかける前の層の出力 $S_c(I)$) を最大にするような入力 I を以下の式を満たすことで作成する手法を提案してる．またこのときネットワークの重みは更新せず，入力のみを更新していく．入力する画像の初期値は $[0, 1]$ の範囲で一様分布から独立にサンプリングされる．

$$\mathbf{I}^* = \arg \max_{\mathbf{I}, s.t. \|\mathbf{I}\|=\rho} S_c(\mathbf{I}) \tag{4}$$

これは，非凸な最適化計画となるが，局所解を見つけることはできる．そこで，Erhan[9] らは，勾配上昇法 (gradient ascent) として以下の更新式を計算することで局所解を求めた．ただし，ϵ は学習率である．

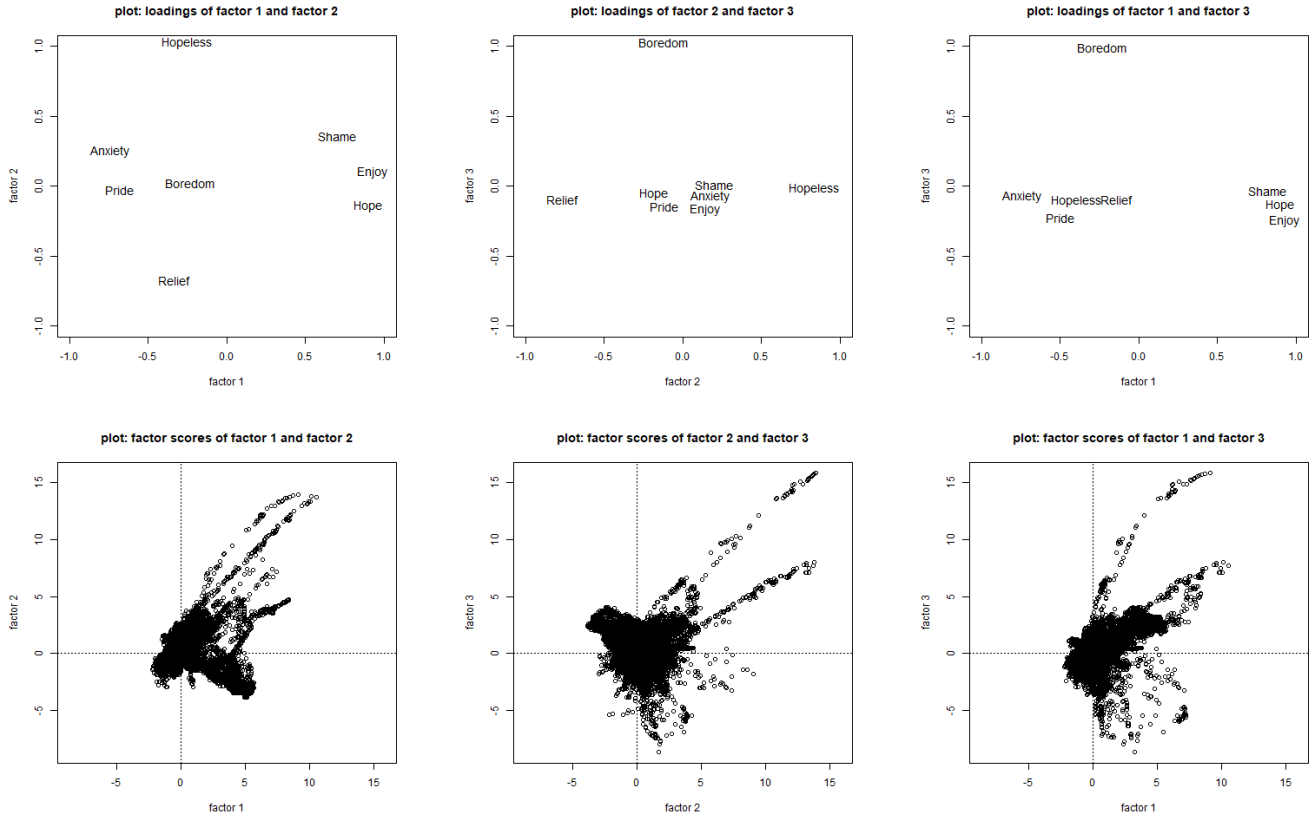

図 3: 各因子の組に関する因子負荷のプロット（上段）と各因子の組に関する各入力データの因子得点のプロット（下段）（左：因子 1 と因子 2，中央：因子 2 と因子 3，右：因子 1 と因子 3）

$$\mathbf{I}_{t+1} = \mathbf{I}_t + \epsilon_1 \frac{\partial S_c(\mathbf{I}_t)}{\partial \mathbf{I}_t} \tag{5}$$

これによって得られた入力は人間にとって解釈が難しい図になることが多く，モデルのデバッグに使うような際には適していない．そこで Simonyan[8] らは下記の式を用いて，入力画像にノルムの制限を加えることで出力を大きくしつつ，人間にとって解釈可能な入力を出力する手法を提案している．

$$\arg\max_{\mathbf{I}} S_c(\mathbf{I}) - \lambda \|\mathbf{I}\|_2^2 \tag{6}$$

5.2 学習器の出力に寄与した入力の可視化

本研究では，Simonyan [8] の手法を採用する．ここで，$\epsilon = 7$，$\lambda = 2$ として学習を行った．ただし今回は，入力の初期値の設定の際に Erhan [9] や Simonyan[8] のように画像データではないことから，データの分布を考慮して，生体情報には標準正規分布からサンプリングを行い，カテゴリカル変数（入力時はダミー変数化）には $[0, 1]$ の一様分布を採用した．また，サンプリングによるバイアスを軽減するためにクラスごとにランダムに 10 回局所解を求めたものを各々単位ベクトルに直したのちに，平均をとっている．以上を踏まえて出力された学習結果を図 4 に示す．入力層のインデックスに関して，「0」は「容積脈波」，「1」は「呼吸の強度」，「2」は「皮膚コンダクタンス」を表しており，3 以降はカテゴリカルデータである教師の発話カテゴリをダミー変数化したものである．濃淡において，色が明るいほど大きい値を，暗いほど小さい値を表している．予測クラスは，0 : Enjoy，1 : Hope，2 : Pride，3 : Anxiety，4 : Shame，5 : Hopeless，6 : Boredom，7 : Relief を表している．

5.3 考察

可視化の結果から，Enjoy，Hope，Pride の予測には，皮膚コンダクタンスが他の要素と比べると特に寄与している傾向がみられる．また Shame，Hopeless，Boredom，Relief に関しては，もっとも寄与している入力のインデックスが発話カテゴリの 9, 10 , 11, 12 でそれぞれ分散的に表現されていることが伺える．この

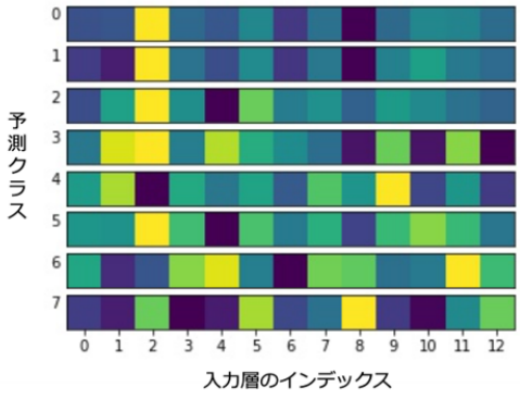

図 4: 各出力が最大になるような入力の単位ベクトルの平均

ことから，Shame，Hopeless，Boredom，Relief に関しては，学習者の生体情報に加え，教師の発話もこれらの心的状態の推定に寄与していることが示唆された．

一方で，今回の可視化のシミュレーションでは，教師発話カテゴリに該当するインデックスに一つ余分なカテゴリを仮定してしまっていることが明らかとなった．この点が可視化に影響を与えてしまっている可能性が考えられるため，このインデックスを除去した学習を行う必要がある．

6　まとめと今後の課題

本研究では，学習者の心的状態を推定する学習器が「何を学習したか」について検討する上で，因子分析手法の適用および Saliency Map による出力に寄与する入力の可視化を試みた．この結果，本機械学習器モデルは，3 つの潜在的な評価観点から学習者の心的状態を推定しており，これらの観点に対する出力への関係性（因子負荷の符号）は，特に学習データに含まれる心的状態の教師データの数に依存していることが示唆され，出力にポジティブに寄与する確信度のようなものに相当することが示唆された．また，どの心的状態のカテゴリにどの入力が寄与しているか，という観点から出力を最大にする入力を可視化することで，Enjoy，Hope，Pride の予測には皮膚コンダクタンスの情報が関与していること，Shame，Hopeless，Boredom，Relief の予測には生体情報に加えて，教師発話も寄与していることが示唆された．

今後の課題としては以下が挙げられる．一点目は，これらの因子分析の結果と Saliency Map の結果がどのように関わっているかについて検討する必要がある．また，本データにおいて，入力カテゴリに 1 つ余分なカテゴリを含めてしまっていた．このカテゴリを除去し，出力に寄与している入力の可視化を引き続き検討する．

参考文献

[1] Matsui, T., Tawatsuji, Y., Fang, S. & Uno, T.: Conceptualization of IMS that Estimates Learners' Mental States from Learners ' Physiological Information Using Deep Neural Network Algorithm. In: Coy A., Hayashi Y., Chang M. (eds) Intelligent Tutoring Systems. ITS 2019. Lecture Notes in Computer Science, vol 11528. Springer, Cham (2019)

[2] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A Survey of Methods for Explaining Black Box Models, *arXiv preprint* arXiv: 1802.01933, pp.1-45 (2018)

[3] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *2017 IEEE International Conference on Computer Vision*, pp.618-626 (2017)

[4] 松居辰則, 宇野達朗, 岡崎桂太, 田和辻可昌: 機械学習を用いた学習者の生体情報からの心的状態推定の試み, 第 42 回教育システム情報学会全国大会 2017 年 8 月 24 日, C4-2 (2017)

[5] 市原 茂: セマンティック・ディファレンシャル法（SD 法）の可能性と今後の課題, 人間工学, 45(5), pp.263-269 (2009)

[6] Keras Visualization Toolkit, `https://github.com/raghakot/keras-vis`, 2019/6/13 閲覧

[7] Nguyen, A., Yosinski,J. & Clune, J.: Understanding Neural Networks via Feature Visualization: A survey, *arXiv preprint* arXiv: 1904.08939, pp.1-23 (2019)

[8] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, *International Conference on Learning Representations Workshop* (2014)

[9] Erhan, D., Bengio, Y., Courville, A. & Vincent, P.: Visualizing Higher-Layer Features of a Deep Network, *University of Montreal* 1341(3), (2009)