学習者の心的状態を推定する機械学習器の解釈可能性を志向した分析的アプローチの提案

Proposal of Analytical Approach for Interpretability of Machine Learning System Estimating Learner's Mental States

田和辻 可昌[*1*2], 古澤 嘉久[*1], 松居 辰則[*2]
Yoshimasa TAWATSUJI[*1*2], Yoshihisa FURUSAWA[*1], 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
Email: y.tawatsuji@aoni.waseda.jp

あらまし: 本研究では，これまで我々が構築してきた学習者の心的状態推定器に関する解釈可能性を志向した分析結果について報告する. 因子分析によって本学習器が獲得した潜在的分類観点を抽出したところ，3 つの潜在的分類観点が抽出された．また，誤った心的状態カテゴリに分類された入力データは因子空間上で一部に局在することが示唆された．一方，Activation Maximization によって出力に寄与する入力ベクトルを可視化した結果，Enjoy，Hope，Pride，Shame の心的状態の推定には教師発話が寄与していることが示唆された．また，生体情報では皮膚コンダクタンスが特に寄与していることが抽出された.
キーワード: Intelligent Mentoring System，機械学習，解釈性，可視化

1. はじめに

近年，教育・学習支援分野において，学習中における学習者の行動履歴や生体情報などの大規模データを収集し，統計的手法を用いて解析するデータアナリティクスが注目されている．我々はこれまで，学習中の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システム，Intelligent Mentoring System（IMS）の構築を進めてきた(1)．一方で，このようなシステムを教育現場へ応用する際には，そのシステムが「何を根拠にその出力を行ったか」という判断根拠の明示, 解釈可能性が重要な課題である．解釈可能性については機械学習の分野で近年活発な議論がなされており，様々な可視化・解釈性の方法が提案されている(2)．ところが，これらは画像を入力としたクラス分類タスクを取り扱うシステムを対象とすることが多く，本課題のような学習ドメインにおける行動履歴や生体情報を入力としたシステムの解釈性は黎明期にあると考えられる．そこで本研究では，学習ドメインデータに基づき学習者の心的状態を推定するシステムの判断根拠の明示を志向し，学習器が「何を学習したか」を多角的な観点から分析・考察することを目的とする．

2 節では，本研究で対象となる学習器システムについて概説する．3 節では，学習器が獲得したと考えられる潜在的分類観点を，因子分析を用いて抽出した結果について述べる．4 節では，機械学習の可視化研究におけるアプローチの一つである Activation Maximization を本学習器に適用した結果について述べ，5 節で今後の課題について述べる．

2. 構築した学習器システム

学習器の構築にあたり，松居ら(3)によって行われた実験から得られた，学習者の容積脈波，皮膚コンダクタンス，呼吸の強度および，教師の発話カテゴリを用いた. 教師データとなる学習者の心的状態は，AEQ（Achievement Emotion Questionnaire）(4)に存在する感情カテゴリに基づいた内省報告を実験後に専用のアプリケーションを通して行ってもらうことで取得した．学習器には中間層 4 層からなる多層ニューラルネットワーク（活性化関数は中間層には ReLU 関数，出力層には Softmax 関数を適用）を採用した．学習には誤差逆伝搬法を用いた．

全データを 6:4 の訓練データと評価データに分割することで学習を行った．学習の結果，学習データにおける Recall（実際に心的状態が X であるデータに対し，機械学習器が「心的状態が X である」と分類したものの割合）は, Enjoy が 52.9%, Hope が 64.6%, Pride が 95.6%，Anxiety が 98.0%，Shame が 70.4%, Hopeless が 76.5%, Boredom が 92.9%, Relief が 99.6% となった．また，検証データにおける Recall はそれぞれ, Enjoy が 50.9%, Hope が 63.6%, Pride が 90.4%, Anxiety が 91.3%, Shame が 68.6%, Hopeless が 72.9%, Boredom が 82.1%，Relief が 94.4%で推定され，本学習器は過学習を起こしていないものと判断した．

3. 因子分析による潜在的評価観点の抽出

機械学習器が獲得した潜在的な分類観点を抽出する方法として因子分析を採用する．因子分析は統計的機械学習の手法の一つであり，データの背後にある要因（共通因子）を推定する手法である．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．平行分析およびカイザー基準の観点から因子数を 3 に決定した．最尤法，プロマックス回転を用いて因子負荷量と共通因子得点を推定した．表 1 に得られた因子負荷行列および各因子の寄与率を示す．

表 1 因子負荷行列および各因子の寄与率

心的状態	因子 1	因子 2	因子 3	共通性
Shame	0.77	0.08	0.04	0.61
Hope	0.76	0.34	-0.13	0.75
Enjoy	0.74	0.1	-0.35	0.68
Hopeless	0.66	0.02	0.14	0.47
Relief	-0.74	0.22	0	0.59
Pride	-0.4	0.44	-0.18	0.46
Anxiety	-0.21	-1.07	-0.31	1
Boredom	-0.05	0.27	1.09	1
寄与率	0.36	0.17	0.16	

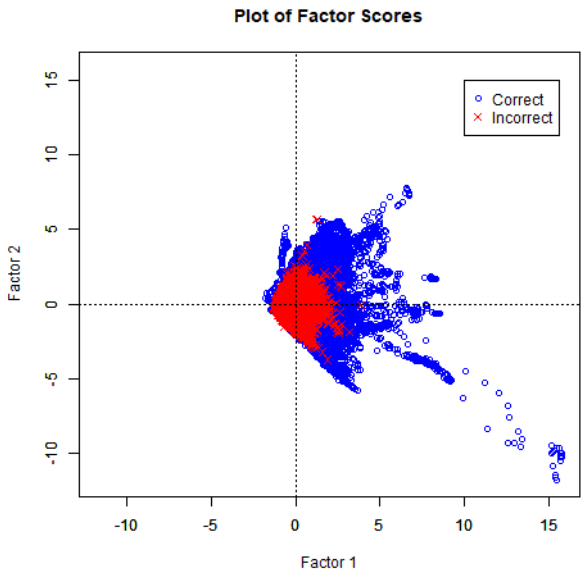

図 1 因子空間における各入力データの配置．青○印は正分類データ，赤×印は誤分類データを表す

この結果から，本機械学習器のモデルは，Shame, Hope, Enjoy, Hopeless, Relief の出力に寄与する第 1 因子，Pride, Anxiety の出力に寄与する第 2 因子，Boredom の出力に寄与する第 3 因子の 3 つの因子を潜在的に有することが示唆された．また，因子空間上に各入力データ（生体情報と教師発話）の因子得点をプロットした結果を図 1 に示す．また，図内の青○印は正しく心的状態を分類できたものを，赤×印は誤って分類したものを示している．この結果から，誤分類されるデータは因子空間上の原点付近に密集していることが示唆される．

4. 出力に寄与する入力表現の抽出

機械学習の解釈性に関わる研究では，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を明らかにすることは重要である．本研究では，Activation Maximization[4][5]を用いて，出力層のユニットを最大にするような入力ベクトルを生成する．具体的には，入力の初期値を[0, 1] の範囲で一様分布から独立にサンプリングし，勾配上昇法によって各出力に最大に寄与する入力ベクトルを求める．各心的状態に対して 10 通りのランダムベクトルを初期値として与え，得られたベクトルの平均ベクトルを可視化したものを図 2 に示す．図の色が明るいほど出力の各カテゴリに対する寄与が高いことを示す．この結果から，Enjoy, Hope, Pride, Shame の分類には，生体情報よりも教師発話が強く影響していることが示唆された．また，生体情報においては，皮膚コンダクタンスが特に出力に寄与していることが示唆された．

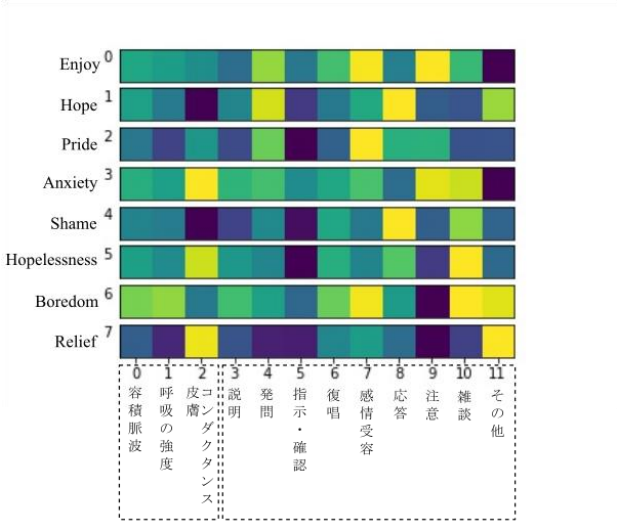

図 2 出力の心的状態（縦軸）に対する入力（横軸）の寄与率（色が明るいほど寄与率が高いことを示す）

5. 今後の課題

今後は，各出力を最大にする入力ベクトルに関して，初期値を変更することで複数用意し，それらが因子空間上でどこにマッピングされるかを検討する．これによって，出力に寄与する入力と因子空間の軸との関係を検討することが可能になると考えられる．

参考文献

(1) Matsui, T. et al.: Conceptualization of IMS that Estimates Learners' Mental States from Learners' Physiological Information Using Deep Neural Network Algorithm, In: Coy, A., Hayashi, Y. and Chang, M. (eds) Intelligent Tutoring Systems. ITS 2019. LNCS Vol. 11528. Springer Cham (2019)
(2) Guidotti, R., et al.: A Survey of Methods for Explaining Black Box Models, arXiv: 1802.01933 (2018)
(3) 松居辰則ら: 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)
(4) Pekrun, R. et al.: Measuring Emotions in Students' Learning and Performance: The Achievement Emotions Questionnaire (AEQ), Contemporary educational psychology. Vol.36, No.1, pp.36-48 (2011)
(5) Erhan, D., Bengio, Y., Courville, A. and Vincent, P.: Visualizing Higher-Layer Features of a Deep Network, University of Montreal, Vol.1341, No.3 (2009)
(6) Simonyan, K., Vedaldi, A. and Zisserman, A.: Deep Inside Convolutional Networks: Visualizing Image Classification Models and Saliency Maps, International Conference on Learning Representations Workshop (2014)