
一般口演A

[KA1] 一般口演 A

機械学習による予測に寄与した特徴量抽出のための相関分析手法を用いた要因分析技術

2018年6月22日(金) 10:50 ～ 11:30 第1会場 (2階・メインホール)

[KA1] 機械学習による予測に寄与した特徴量抽出のための相関分析手法を用いた要因分析技術

鈴木 麻由美（株式会社日立製作所 研究開発グループ）

機械学習による予測に寄与した特徴量抽出のための 相関分析手法を用いた要因分析技術

鈴木 麻由美^{*1}, 柴原 琢磨^{*1}, 村垣 善治^{*2}

^{*1} 日立製作所 研究開発グループ

^{*2} 東京女子医科大学 先端生命医科学研究所 先端工学外科学分野

Factor Analysis Technique with Correlation Analysis Method to Extract Feature Contributed to Prediction by Machine Learning

Mayumi Suzuki^{*1}, Takuma Shibahara^{*1}, Yoshihiro Muragaki^{*2}

^{*1} Hitachi, Ltd. Research and Development Group

^{*2} Faculty of Advanced Techno-Surgery, Institute of Advanced Biomedical
Engineering and Science, Tokyo Women's Medical University

抄録:

SVM や deep neural network などの新しい機械学習手法により予測精度は向上したが、これらの手法は非線形な機械学習モデルを利用するため、予測の根拠に関する説明性に欠ける。予測の根拠の説明性が向上することにより、予測結果の信頼性も向上する。そこで、非線形な機械学習モデルにおいても、予測に利用された特徴量の提示を可能にする要因分析技術を提案した。要因分析技術では、逆解析技術により算出した機械学習モデルの事後確率分布に対し、因子解析技術により非線形な相関関係の評価を行うことにより、アウトカムとの相関が強い特徴量を抽出する。前立腺腫瘍患者と健常者の遺伝子発現データを用いた実験を行い、前立腺腫瘍や細胞死などとの関連性が指摘されている遺伝子を抽出した。今後、因子解析技術が抽出する特徴量の特性の検証や、本実験により抽出された遺伝子を用いた臨床研究が期待される。

キーワード 機械学習, ディープラーニング, 因子解析, ゲノミクス

1. はじめに

近年、deep neural network(DNN)をはじめとする非線形な機械学習モデルにより、画像および音声認識の分野において、高精度な予測が可能になりつつある。ヘルスケア分野においても、非線形な機械学習モデルを用いた疾病や薬剤の有害事象などの解析が行われ始めている。一般に、非線形な機械学習モデルを用いた場合、高い予測精度を得られる可能性があるが、モデルの構造が複雑なため、予測に利用された特徴量を提示することは不可能である。そこで、非線形な機械学習モデルにおいても予測に利用された特徴量を提示する要因分析技術を提案した[1]。予測精度に加えて、予測に利用された特徴量を用いた予測結果の信頼性の評価が可能となった。

提案した要因分析技術は、以下の2つの技術から構成した。まず、逆解析技術を用いて、機械学習モデルをベイズ統計により事後的に解析する。次に、因子解析技術を用いて、算出した事後

確率分布に Jensen-Shannon divergence(JS divergence)を適用し、アウトカムにおいて事後確率分布の差異が大きい特徴量を抽出する。

本研究では、要因分析技術における予測に利用された特徴量の探索能力の向上を目的として、非線形な相関関係の評価可能である Backward Elimination Using Hilbert-Schmidt Independence Criteria(BAHSIC)を因子解析技術として試した結果について報告する。

2. 方法

1) 相関係数を用いた因子解析技術

逆解析技術により得られた事後確率分布に対し、BAHSICを用いて2つの量的変数間の相関が強い特徴量を抽出する。BAHSICは、データを相関係数が最大となる高次元の再生核ヒルベルト空間に写像することにより2つの変数間の相関関係を算出する手法であり、非線形な相関に関しても算出可能である[2]。特徴量を x 、アウトカムを y 、 x および y の可分な再生核ヒルベルト空間を $\mathcal{F}$ お

および $\mathcal{G}$, 結合確率測度を p_{xy} , 交差共分散行列を C_{xy} とすると, HSICは C_{xy} の平方ヒルベルトシュミットノルムとして式(1)で定義される.

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|_{\mathcal{HS}} \quad (1)$$

式(1)が 0 である場合, x と y が独立, すなわち無相関であることを示す. BAHSIC では, 全特微量とアウトカムの独立性と, 一部の特微量とアウトカムの独立性を比較し, 独立性に影響を及ぼさない特微量の削除を繰り返すことにより, 削除された順に独立性が低い特微量として順位をつける.

2) 実験方法

分析データとして, Singh の論文より取得した前立腺腫瘍患者と健常者の遺伝子発現データを用いた[3]. 遺伝子発現データは, 特微量数が 1,000, サンプル数が 92 の臨床データである. 機械学習モデルは DNN とし, 交差検証により最適なモデルを決定した. なお, DNN の予測性能の比較のため, 多項式カーネルを用いた SVM による予測を行った.

3. 結果

まず, DNN により構築した機械学習モデルの予測精度を Table. 1 に示す. Accuracy が 0.977, Precision が 0.966, Recall が 0.983, F-Measure が 0.960 となった. DNN を用いたことで, SVM による予測精度より 5%程度の改善を示した. 次に, 本モデルに要因分析技術を適用した結果, 第 1 位は VDAC1, 第 2 位は CYP2C8, 第 3 位は PHIP, 第 4 位は TRPM2, 第 5 位は ANGEL2 であった. これら 5 個の遺伝子は, 論文中に重要性について報告はないが, VDAC1 と ANGEL2 は前立腺腫瘍, CYP2C8 は肝臓腫瘍, PHIP はメラノーマ, TRPM2 は前立腺の細胞死との関連性が指摘されている遺伝子であった. また, JS divergenceを用いた既報告と BAHSICを用いた本報告における抽出した特微量の上位 5 位, 10 位, 100 位に含まれる因子の一致率はそれぞれ, 40%, 40%, 49%であった.

Table. 1. 予測精度

	Accuracy	Precision	Recall	F-Measure
DNN	0.977	0.966	0.983	0.960
SVM	0.924	0.947	0.915	0.925

4. 考察

NEBL は, JS divergence および BAHSIC により抽

出された上位 5 位のうち, 前立腺や腫瘍との関連性が唯一指摘されていない. NEBL は, 拡張性心筋症との関連性が指摘されている. 拡張性心筋症と前立腺腫瘍は, 高齢の男性の罹患率が高いなど患者背景が類似している. そのため, 拡張性心筋症と前立腺腫瘍の分布は類似していると考えられる. 一方で, BAHSIC で抽出されないことから, 拡張性心筋症の患者と健常者の分布において, 一部分では差異が認められるが, 前立腺腫瘍ではないため一部分では差異が認められないため, 全体として低い相関となっていると考えられる. このような分布となる特微量は一部の患者にのみ特異的な特微量や, 分布間の一部の類似性からこれまで注目されなかった特微量が該当すると考える. つまり, JS divergenceはこれまでの統計的解析では注目されなかった特微量を抽出可能であると考えられる.

以上から, 複数の因子解析技術を用いることで, 要因分析技術における予測に利用された特微量の探索能力が向上した. 今後, 各技術の抽出特性を検証するためには, データセットによるベンチマークを行う必要がある. また, 本実験において提示された特微量における臨床的な腫瘍との関連性の調査には, 前向き研究が必要となる.

5. 結語

BAHSIC を用いた因子解析技術により, 非線形な機械学習モデルにおいて予測に利用された特微量を新たに提示できた. 今後, 抽出した特微量を用いた臨床研究への応用が期待される.

参考文献

- [1] 鈴木麻由美, 柴原琢磨, 村垣善治: 機械学習による予測に寄与した特微量抽出のための要因分析技術, JCMI, vol.37 pp. 854-856, Nov. 2017.
- [2] Gretton, Arthur, Olivier Bousquet, Alex Smola, et al: Measuring statistical dependence with Hilbert-Schmidt norms, Algorithmic learning theory. Springer Berlin Heidelberg, 2005.
- [3] D. Singh, P. G. Febbo, K. Ross, et al: Gene expression correlates of clinical prostate cancer behavior, Cancer cell, vol. 1, no. 2, pp. 203-209, 2002.