一般口演

一般口演28
機械学習・アルゴリズム・解析モデル

2017年11月23日(木) 12:45 ～ 14:15 E会場 (10F 会議室1003)

[4-E-2-OP28-1] 機械学習による予測に寄与した特徴量抽出のための要因分析技術

鈴木 麻由美[1], 柴原 琢磨[1], 村垣 善浩[2] （1.株式会社 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）

【目的】
近年、 deep learningをはじめとする非線形な機械学習モデルにより、高精度な予測が可能になりつつある。しかし、非線形な機械学習モデルではモデルの構造が複雑なため、 logistic regressionに代表される線形な機械学習モデルのようにパラメータから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する。
【方法】
要因分析技術は、機械学習モデルをベイズ統計により事後的に解析する逆解析技術と、逆解析技術により算出した事後確率分布を Jensen Shannon divergenceを用いて確率分布同士の差異が大きい特徴量を重要な特徴量として抽出する因子解析技術から構成した。機械学習モデルは deep learningとし、交差検証により最適なモデルを決定した。なお、人工データおよび、 Singhら(2002)による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に対し要因分析技術を適用した。
【結果】
人工データにおいて、要因分析技術による抽出結果の上位10個が予め設定した予測に寄与する特徴量10個と一致した。遺伝子発現データにおいて、抽出結果の上位5個のうち4個は、論文では重要と報告されていないものの、前立腺以外の腫瘍との関連性が指摘されている遺伝子であった。
【考察】
遺伝子発現データの結果から、線形な機械学習モデルを用いた解析では発見できなかった特徴量を提示できたと考えられる。但し、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。
【結論】
要因分析技術により、非線形な機械学習モデルにおいて予測に利用された特徴量を提示可能であることが確認できた。今後、抽出された重要な特徴量から新たな臨床研究への応用が期待される。

機械学習による予測に寄与した特徴量抽出のための要因分析技術

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 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

Jumping five decades forward to the present, we now have a number of new learning methods including SVMs, neural networks, and deep learning that are very accurate, but unfortunately also the lack of explainability. Especially, deep learning provides no information about the importance of feature variables. Addressing this problem, we developed a factor analysis technique for the nonlinear machine learning methods. The technique has two statistical steps as follows. The first step, called backward analysis, generates the probability distribution of the positive and negative classes that the prediction model estimated. The second step extracts features that have the significant difference between the positive and negative class distributions by using Jensen Shannon divergence. The factor analysis technique was verified by simulation. Through the experiment, we extracted new factors, have relevance to prostate cancer, from the features of gene expression data. Experimental results show the great potential of our technique to play a vital role in the clinical research.

Keywords: Machine Learning, Deep Learning, Factor Analysis, Genomics

1. 背景および目的

近年, deep learning 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。ヘルスケアの分野においても，非線形な機械学習モデルを用いたデータ解析による疾病や薬剤の有害事象などを解析する試みが行われ始めている。これまで主に使われてきた logistic regression に代表される線形な機械学習モデルでは，達成可能な精度に原理的な制約があったが，非線形な機械学習モデルではより高い精度を得られる可能性がある[1]。反対に，線形な機械学習モデルでは，パラメータから予測に利用された特徴量を提示することが可能であるが，非線形な機械学習モデルでは，モデルの構造が複雑なためパラメータから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する。

2. 方法

2.1 機械学習の特徴

まず, 機械学習の特徴について説明する。代表的な機械学習手法には，logistic regression（LR），support vector machine（SVM），deep learning（DL）などがある。線形な機械学習モデルである LR では，入力データを学習し，重みベクトルとバイアス項を決定する。出力を y, 特徴量が D 次元の入力データを $\boldsymbol{x}$ とすると，出力 y は式(1)で与えられる。

$$y = \sigma((1+\exp(-(\mathbf{w}^T\mathbf{x}+\mathbf{b})))) \quad (1)$$

ここで，$\boldsymbol{w}$ は D 次元の重みベクトル，$\boldsymbol{b}$ は D 次元のバイアス項，関数 σ は sigmoid 関数である。式(1)より，重みベクトル $\boldsymbol{w}$ の値が大きい特徴量ほど予測結果に影響を与える特徴量であることがわかる。つまり，各特徴量における重みベクトル $\boldsymbol{w}$ の値が大きい特徴量が，予測に利用された特徴量であると判断可能である。

次に，非線形な機械学習モデルである DL では，入力データを学習し，各層における重みベクトルとバイアス項を決定する。1 層目の出力を $\boldsymbol{h_2}$，特徴量が D 次元の入力データを $\boldsymbol{x}$ とすると，1 層目の出力である $\boldsymbol{h_2}$ は式(2)で与えられる。

$$\mathbf{h}_2 = f(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1) \quad (2)$$

ここで，$\boldsymbol{W_1}$ は 1 層目における重み行列，$\boldsymbol{b_1}$ は 1 層目におけるバイアス項，関数 f はアクチベーション関数である。アクチベーション関数とは，中間層で用いられる関数であり，sigmoid 関数や Rectified Linear Unit（ReLU）関数などが用いられる[2][3]。

続いて，最終層の出力を y，最後の中間層である M-1 層目の出力を $\boldsymbol{h_M}$ とすると，出力 y は式(3)で与えられる。

$$y = \sigma(\mathbf{W}_\mathbf{M}\mathbf{h}_\mathbf{M}+\mathbf{b}_\mathbf{M}) \quad (3)$$

ここで，$\boldsymbol{W_M}$ は最終層である M 層目の重み行列，$\boldsymbol{b_M}$ は M 層目のバイアス項である。式(2)および式(3)から，入力データ $\boldsymbol{x}$ は，各層における計算により異なるデータに変換され，最終層まで伝搬していることがわかる。そのため，入力データ $\boldsymbol{x}$ における各特徴量が予測結果にどのような影響を与えているか解析することは困難である。

そこで，本研究では，非線形の機械学習モデルから予測に利用された特徴量を抽出する要因分析技術を提案する。要因分析技術は，機械学習モデルをベイズ統計により事後的に解析する逆解析技術と，逆解析技術により算出した事後確率分布に Jensen-Shannon divergence（JS divergence）を適用することで確率分布同士の差異が大きい特徴量を重要な特徴量として抽出する因子解析技術から構成した。以下，2 つの技術について説明する。

2.2 要因分析技術

2.2.1 逆解析技術

まず機械学習モデルを構築する。例えば，がん患者と健常者のデータを学習させ，モデルを構築する。構築した機械学習モデルに，新たに患者のデータを入力すると，がん患者かどうかを予測することが可能となる。

次に，構築した機械学習モデルに対して，Expanded Ensemble Monte Carlo 法の1つであるレプリカ交換モンテカルロサンプリング法(Replica Exchange Markov Chain Monte Carlo Methods, RMC) を用いて，がん患者である確率が最大，もしくは，最小となる因子の分布を効率的に導出する．RMC は，メトロポリスヘイスティング法（Metropolis-Hasting, MH）のアルゴリズムをベースとしており，モンテカルロ法（Monte Carlo method）やマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo methods）のサンプリング効率を改善した方法である。RMC は，異なる初期値から始まる N 個の系を用いて MH 法の基準でサンプリングを行う。異なる初期値から始まる N 個の系を用いることで，初期値の影響を受けにくくなる。MH 法によるサンプリングでは，目標分布を分散 σ の Boltzmann 分布として与え，採択確率と一様分布に従いサンプリング値を得る。なお，初期値の影響を避けるため，通例，サンプリング開始から数十%のサンプルを除去する。この除去期間を burn-out 期間と定める。以上の処理により，がん患者である確率に対する事後確率分布を算出することが可能となる。

RMC によりサンプリングされた結果の収束性は，Gelman-Rubin 診断を用いた。Potential scale reduction factor（PSRF）$\hat{R}$ が 1.0 に近づくほど良く，1.1 から 1.05 程度になれば，収束したと判断可能となる。本研究では，異なる初期ベクトルから計算された 2 つの特徴量ベクトルを小数点以下第 3 位まで評価した[4]。

2.2.2 因子解析技術

単独の特徴量として予測に利用される特徴量とは，がん患者と非がん患者における各特徴量を比較した際に，値の分布が異なる特徴量であると考えられる。そこで，逆解析技術により得られた事後確率分布において，JS divergence を用いて，がん患者である事後確率分布とがん患者でない事後確率分布を比較する。

まず，逆解析技術により，がん患者である確率が高い患者の確率分布 $P(x_n \mid y > 0.8)$と確率が低い患者の確率分布 $P(x_n \mid y < 0.8)$を導出する。ここで，x_n は n 番目の特徴量の分布，y はがん患者である確率値である。次に，2 つの確率分布に対して各特徴量に，JS divergence を適用し，値が大きい順に特徴量を並べる。以上の処理により，予測に利用された特徴量として重要な順に得ることができる。

2.3 実験方法

分析データとして，人工データおよび，Singh らによる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[5]。人工データは，特徴量数およびサンプル数を 1000，重要な特徴量数を 10 として生成した人工データである。重要な特徴量は，正例および負例ごとに平均値をずらしたカイ 2 乗分布として生成し，重要でない特徴量は，一様分布から生成した乱数とした。遺伝子発現データは，特徴量数は 1000，サンプル数は92の臨床データである。患者サンプルである正例は 47 個，患者ではないサンプルである負例は 45 個であった。一般的に，各特徴量間では単位が異なるため，値の絶対値に大きな差が生じている場合がある。この場合，予測モデルを構築する際に，値が大きい数字に結果が依存した局所解に陥る危険性がある。そのため，各データは各特徴量について正規化を行った。

機械学習モデルは deep learning とし，交差検証により最適なモデルを決定した。ハイパーパラメータの探索範囲を表 1 に示す。各層のアクチベーション関数の種類などについては，表 1 に示す範囲のなかで Bergstra らの手法を用いてランダム探索ベースの最適化を実施した[6]。また，計算コストと過学習の抑制のため，drop-out を使用した[7]。

表 1 ハイパーパラメータの探索範囲

Hyper parameter	Optimization range
Number of layers	4, 5
Activation function	Sigmoid, tanh, ReLU
Number of hidden units	100 to 500
Dropout rate	0.1 to 0.5
Regularization function	L1, L2 or Elastic

逆解析技術の収束性は，Gelman-Rubin 診断により $\hat{R}$ の値により判断した。要因分析技術のパラメータは，σ を $0.1<\sigma<0.5$ の定義域で均等に，10 本の独立連鎖を生成して，それぞれのサンプリング回数を 100,000 回とした。また，burn-out 期間として，前半 50 %のサンプルをデータから除外し，$0.25<\sigma$ の独立連鎖 5 本を経験的事後確率分布として採用した。

3. 実験結果

3.1 人工データ

人工データを用いた実験結果を示す。要因分析技術による抽出結果の上位 10 個が，データ生成の段階にて設定した予測に寄与する特徴量 10 個と一致した。なお，各特徴量に重要度の順位は定めていないため，抽出結果の順序に関係はない。また，逆解析技術の収束性を示す $\hat{R}$ の値は収束したことを示す 1.00 となった。

3.2 遺伝子発現データ

遺伝子発現データを用いた実験結果を示す。まず，構築した機械学習モデルのハイパーパラメータを表 2 に示す。また，予測精度は，Accuracy が 0.977, Precision が 0.966, Recall が 0.983, F1 が 0.960 となった。本モデルに対し，要因分析技術を適用した結果を示す。解析技術の収束性を示す $\hat{R}$ の値は収束したことを示す 1.00 となった。抽出した予測に利用された特徴量のうち，上位 5 位の特徴量を表 3 に示す。第 1 位は TIAL1, 第 2 位は PHIP, 第 3 位は PCNA, 第 4 位は VDAC1, 第 5 位は NEBL であった。

表 2 ハイパーパラメータ

Hyper parameter	Optimization range
Number of layers	4
Activation function	ReLU
Number of hidden units	143
Dropout rate	0.3
Regularization function	L1

表 3　上位 5 位の予測に利用された特徴量の一覧

Ranking	Gene Symbol	Gene Title
1	TIAL1	TIA1 cytotoxic granule-associated RNA binding protein-like 1
2	PHIP	Pleckstrin homology domain interacting protein
3	PCNA	proliferating cell nuclear antigen
4	VDAC1	voltage-dependent anion channel 1
5	NEBL	nebulette

4. 考察

人工データを用いた実験において，予測に利用された特徴量として提示した上位 10 個の特徴量は，予め設定した予測に寄与する特徴量 10 個と一致したことから，要因分析技術により，正しく予測に利用された特徴量を抽出可能であることを確認できた。

遺伝子発現データを用いた実験において，予測に利用された特徴量として提示した特徴量は，TPA や PSA，p53 のような既に前立腺腫瘍の発現に関与すると示されている遺伝子ではないため，これまでの線形な機械学習モデルを用いた解析では発見できなかった特徴量を提示できたと考えられる。しかし，真に発現に関与しているかどうかは不明である。まず，1 位であった TIAL1 は，細胞毒性関連蛋白質であり，細胞傷害性 T 細胞(cytotoxic T lymphocyte, CTL) に発現し，腫瘍への関連性があると指摘 されている遺伝子である。 次に，2 位であった PHIP は，メラノーマ（悪性黒色腫）の識別をする際の有用性が指摘されている腫瘍関連因子である。3 位であった PCNA は，腫瘍細胞の生物学的悪性度や増殖能の判定および悪性腫瘍の治療方針，予後の検討に極めて有用な因子である。4位であった VDAC1 は，前立腺癌細胞におけるアポトーシスの外因性経路の正の調整因子 であることが指摘されている遺伝子である。最後に，5 位であった NEBL は，拡張型心筋症などに関連する因子であり，腫瘍との関連性に関しては不明な遺伝子である。上位 5 個のうち 4 個は，論文では重要と報告されていないものの，前立腺以外の腫瘍との関連性が指摘されている遺伝子であった。そのため，本実験において提示された特徴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. 結論

要因分析技術により，非線形な機械学習モデルにおいて予測に利用された特徴量を提示可能であることが確認できた。今後，抽出された重要な特徴量から新たな臨床研究への応用が期待される。

参考文献

[1] M. Minsky and S. Papert, “Perceptrons.” 1969.

[2] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models”, Proc. ICML, vol. 30, pp. 1, 2013.

[3] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines”, Proc. the 27th International Conference on Machine Learning (ICML-10), pp. 807–814, 2010.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Bayesian data analysis. Taylor & Francis, vol. 2, 2014.

[5] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, J. P. Richie et al., “Gene expression correlates of clinical prostate cancer behavior”, Cancer cell, vol. 1, no. 2, pp. 203–209, 2002.

[6] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization”, The Journal of Machine Learning Research, vol. 13, no. 1, pp. 281–305, 2012.

[7] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors”, arXiv preprint arXiv:1207.0580, 2012.