

一般口演 | 医療データ解析

一般口演11 機械学習

2021年11月20日(土) 09:10 ~ 11:10 G会場 (2号館3階232+233)

[3-G-1-01] 蛋白分画検査の波形情報を使用した血液検査の欠測値補間の試み

*兵頭 勇己¹、久原 太助²、檜山 麻里子¹、安井 繁宏¹、畠山 豊¹、奥原 義保¹（1. 高知大学医学部附属医学情報センター, 2. 高知大学医学部附属病院検査部）

*Yuki Hyohdoh¹, Taisuke Hisahara², Mariko Hiyama¹, Shigehiro Yasui¹, Yutaka Hatakeyama¹, Yoshiyasu Okuhara¹（1. 高知大学医学部附属医学情報センター, 2. 高知大学医学部附属病院検査部）

キーワード：serum protein electrophoresis, blood chemical analysis, machine learning, missing data

【背景】

血液検査結果は Real World Data study(RWDs)で多用されるが、様々な要因にて欠測が発生する。本研究では、蛋白分画検査結果のうち、分析精度が高いキャピラリー電気泳動法による波形データを使用し、種々の血液検査結果を識別するためのモデルを網羅的に構築、欠測値の補間手段となり得るか検討する。

【方法】

高知大学医学部附属病院にて蛋白分画検査を実施した20歳以上の患者（2010年～2015年）を対象とし、病院情報システムに蓄積されたデータを使用した。波形データはプレアルブミン分画側からγグロブリン分画側までの計300区画の易動度を使用し、総蛋白値による補正を実施した。対象の検査項目は、CBCや生化学検査など頻度の高い22項目とし、蛋白分画検査と同日に実施した結果値を使用した。年齢および性別、波形データを説明変数、目的変数を対象検査項目の結果値を基準範囲内か否かとした識別モデルを作成し、モデル候補として決定木系およびニューラルネットワーク系のアルゴリズムを網羅的に作成した。対象データから独立した検証用データを400件、残りの75%をモデル構築用、25%をパラメータ調整用データとしてランダムに抽出し、モデルの評価を AUC値にて実施した。

【結果】

対象の波形データは2379件であった。多くの検査項目において勾配ブースティング決定木が識別性能の優位性を示し、中でも C反応性蛋白に関しては AUC値0.90と、既存手法であるロジスティック回帰モデルよりも良好であった。

【考察】

各種疾患に対する波形パターンが既知であり、かつ血液検査結果に反映される検査項目に関しては少なくとも精度よく識別可能であることが示唆された。本手法を展開することで、安価であるが臨床的に活用頻度が低い蛋白電気泳動検査の波形データを、RWDsで問題となる欠測値の補間手段の1つとして活用できると考える。

蛋白分画検査の波形情報を使用した血液検査の欠測値補間の試み

兵頭勇己^{*1}、久原太助^{*2}、檜山麻里子^{*1}、安井繁宏^{*1}、畠山豊^{*1}、奥原義保^{*1}

*1 高知大学医学部附属医学情報センター、*2 高知大学医学部附属病院医療技術部臨床検査部門

An Interpolation of Missing Values in Blood Tests Using Serum Protein Electrophoresis Diagram

Yuki Hyohdoh^{*1}, Taisuke Hisahara^{*2}, Mariko Hiyama^{*1}, Shigehiro Yasui^{*1},
Yutaka Hatakeyama^{*1}, Yoshiyasu Okuhara^{*1}

*1 Center of Medical Information Science, Kochi Medical School, Kochi University,

*2 Department of Clinical Laboratory, Kochi Medical School Hospital, Kochi University

Blood test results are frequently used in epidemiological studies, however missing data can occur due to various factors. In this study, we developed a machine learning model for predicting blood test results using serum protein electrophoresis diagram. The purpose of this study was to evaluate whether the prediction by the machine learning model can be used to interpolate missing values. Patients over 20 years old who test serum protein electrophoresis at Kochi Medical School Hospital were included in the study. Using serum protein electrophoresis diagram, we generated a machine learning model to identify whether the results of 21 test items were within the reference intervals and compared it with logistic regression model. In addition, pseudo-missing data sets were generated, and the variability of the estimated parameters were compared between the interpolation method by our machine learning model and multiple imputation method. Good performances were obtained with AUC values of 0.9 or higher for some of the blood tests. Compared to the parameter estimation by the multiple imputation method, the interpolation by the machine learning model showed less variation in parameters. The missing value interpolation using serum protein electrophoresis diagram is a practical method.

Keywords: serum protein electrophoresis, blood chemical analysis, machine learning, missing data

1. 結論

血液検査は、臨床現場において患者に対して実施する検査のうち、頻度の多い検査の一つである。血液検査の結果は、臨床医が行う臨床推論に大きな影響を与え、患者の治療方針やケアに重要な影響を及ぼす。また、病院情報システムに蓄積された検査結果のデータは、単一施設もしくは多施設共同研究などの疫学研究で利用されることも多い。しかしながら、病院情報システムに蓄積された検査結果のデータは診療および医療保険制度などの理由にて欠測が多く発生する。

蓄積された病院情報システムのデータ活用にあたり、欠測値の対処は必須であるが、その解決は難しい。特にデータの発生が診療上の必要性に依存する検査データなどは、欠測がランダムに発生する MAR (Missing At Random) や MCAR (Missing Completely At Random) ではなく、MNAR (Missing Not At Random) に分類される。MNAR では欠測値代入法として利用される多重代入法では一般に推定値にバイアスが生じ、適切ではない。独立した他の情報から欠測値そのものを精度よく推定することができれば、検査データの欠測値への対処の一つとなりうる。

蛋白分画検査は、容易に臨床検査室にて実施可能であり、かつ安価な検査である一方、最近ではその検査数が減少傾向にある。しかしながら、蛋白分画検査にて得られる波形情報は幅広いスペクトルを持つことから、特定の検査項目においては波形情報から血液検査値を推測できる可能性が示されており、欠測値の推定に利用できる可能性がある。しかし、疫学研究で多用される血液検査項目に対する予測性能や、波形情報を用いて実際のデータ欠測プロセスに対処できるかどうかは明らかになっていない。

2. 目的

本研究は、スクリーニング検査項目を対象に、蛋白分画検査の波形情報から検査結果の予測を実施するための堅牢な予測モデルを構築すること、さらに擬似的に欠測値を発生させたデータセットを使用し、欠測値の補間に対して構築した予測モデルの利用可能性を検証することを目的とした。

3. 方法

3.1 解析対象データ

高知大学医学部附属病院にて蛋白分画検査を実施した20歳以上の患者(2010年～2015年)を対象患者とし、すべてのデータは病院情報システムに蓄積されたデータを使用した。

解析対象データは、蛋白分画検査の波形情報、蛋白分画検査と同日に実施した血液検査結果、および患者基本情報(年齢・性別)とした。血液検査項目は、血球数算定や生化学検査などの検査実施頻度が高い21項目とした。

3.2 データ処理

キャピラリー電気泳動法による波形情報はアルブミン分画から $\alpha 1$ 分画の間のデータを含む、プレアルブミン分画側(陽極側)から γ グロブリン分画側(陰極側)までの計300区画の易動度を使用した。波形情報の補正および正規化は、アルブミンのピーク位置を易動度75となるよう調整し、さらに、総蛋白値が7.0g/dLの時、曲線下面積が100000となるように波形を補正した。

血液検査項目の結果値は、年齢および性別に応じて病院内で設定されている基準範囲を閾値とし、「基準範囲内」「範囲外」の二値データに変換した。

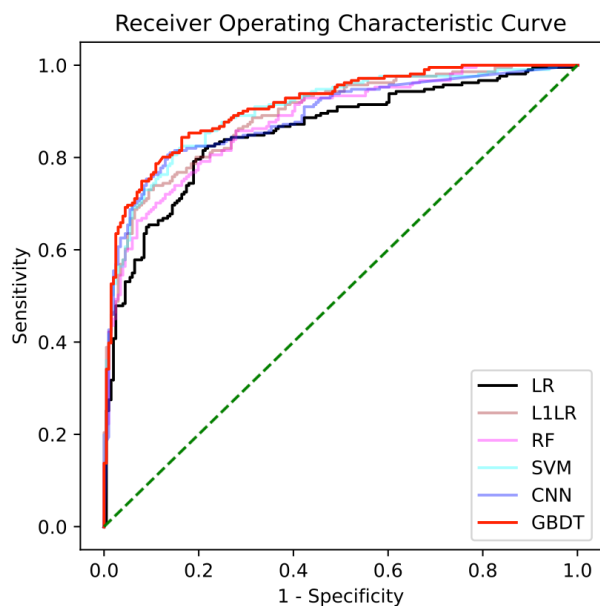


図2 C反応性蛋白の検査結果予測における各機械学習モデルのROC曲線

目は10項目であった。また、6つの学習モデルの中で、GBDTモデルはAUC値の最高値を示すことが多く、検査21項目中、12項目においてGBDTモデルが最良モデルであった。検査項目の例としてC反応性蛋白の予測における、各学習モデルのROC曲線を図2に示す。ROC曲線においても、既存手法であるLRモデルよりも、各種学習モデルの方が識別性能が高いことが確認された。

疑似欠測データセットを使用し、欠測値に対処する3通りの手法による、推定された総コレステロールのオッズ比の分布の違いを示した図を図3に示す。蛋白分画検査の波形情報から欠測値を予測する手法においては、リストワイズ法と多重代入法による欠測値対処法と比較し、推定されたオッズ比の中央値は、欠測の割合に関わらず完全データから推定された真のオッズ比から大きな偏りは認められなかった。また、欠測の割合が増加するほど、リストワイズ法および多重代入法による欠測値対処法では推定値のばらつきが増大する傾向が認められた。

5. 考察

本研究の目的は、蛋白分画検査の波形情報から検査結果の予測を行うための機械学習モデルを構築することであり、さらに欠測値の補間に対する有効性を評価することであった。実際、すべての対象検査項目において、既存の手法であるLRモデルよりも良い識別性能を示す機械学習モデルを見つけ出すことができた。さらに、疑似欠測データセットを使用した評価において、蛋白分画検査の波形情報から欠測値の対処を実施すると、推定されるオッズ比のばらつきが小さくなる傾向が認められた。

蛋白分画検査は、多くの情報が得られる有用な検査であり、血清中には電気泳動でのみ検出可能な異常蛋白も存在し¹⁾、蛋白の種類によっては鋭敏に波形に反映する領域が存在する。また、いくつかの疾患には波形パターンと臨床的意義が明らかになっている。本結果では、それぞれの分画領域で表現される炎症系のマーカーや、貧血関連の指標においては概ねAUC値が0.8~0.9以上となる、妥当な結果を得ることがで

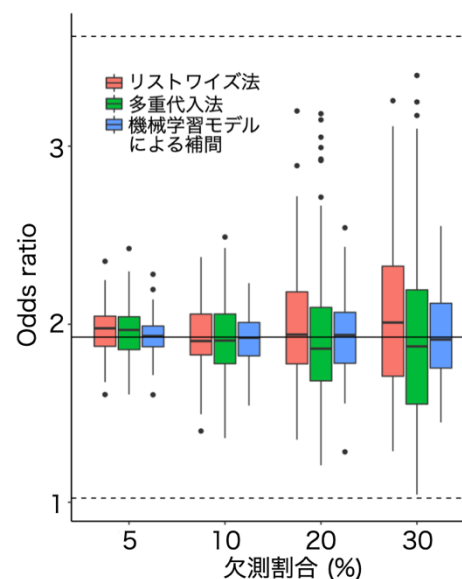


図3 疑似欠測データセットにて推定した総コレステロール値のオッズ比の分布。赤色はリストワイズ法、緑色は多重代入法、青色は波形情報から欠測値を予測し補間した結果を示す。実線は完全データのオッズ比の推定値、点線は95%信頼区間を示す。

きた。さらに、他の項目においても高いAUC値を得ている項目が存在しており、これは検査項目そのものの物質の検出と合わせて、他の蛋白関連の指標と連動することで波形情報に反映されたものをモデルが学習した結果であると考ええる。

ニューラルネットワーク系のアルゴリズム、特に次元量込みニューラルネットワークよりも、決定木をベースとした学習モデルの方が全体的に良い性能となった。これは、パラメータ調整を含めて実際に学習データとして使用できた件数が、1700件前後であり、パラメータ数の多いCNNでは十分に学習ができなかった可能性がある。また、波形の前処理を適切に実施し、テーブルデータ形式にて波形を表現していたことから、決定木系のアルゴリズムであるGBDTの優位性が表れたと考える。検査項目によっては、同じくニューラルネットワーク系の手法であるSVMモデルの方が学習モデルとして高性能であったため、事前モデルとして、波形の特徴抽出に特化して学習させた次元CNNモデルなどを転移学習することで性能が向上する可能性はある。しかし、ネットワークの構成検討やパラメータ探索などの総合的な学習時間を考慮すると、蛋白分画検査の波形情報を識別するためのモデルとしてはGBDTなどの決定木系の手法が最も実用的、かつ十分と考える。

疑似欠測データセットに対して、機械学習モデルから欠測値を補間した場合、推定されるパラメータのばらつきが他の手法と比較して小さくなることが確認された。これは、MNARを仮定した欠測プロセスの場合、蛋白分画検査の波形情報という独立した情報を使用した欠測値の対処が有効である可能性を示している。本研究に限らず、後方視的な疫学研究を実施する場合、血液検査の結果値は利用されることが多い。しかし、検査オーダはあくまで臨床医が患者の診療上の必要性に応じて実施するものであり、研究目的で収集されたデータとは異なり多数の欠測値が存在することとなる。Weberらは、血液検査の実施頻度の情報から検査の正常範囲が推測でき

る可能性を示した²⁾。また、Zhangらは、集中治療室入室患者を対象とし、検査の実施回数は独立して死亡率と関連していることを示した³⁾。これらは、患者の状態、すなわち検査結果値そのものが検査の実施頻度と関係しており、血液検査の欠測プロセスはその値に依存する、MNARであることを示している。一般に、MNARに対して欠測データを除去するリストワイズ法や、他の変数を元に欠測値を補間する多重代入法などでは、推定値のバイアスを除去できないとされる⁴⁾⁵⁾。蛋白分画検査の波形情報を使用した欠測値補間は、MNARを仮定した擬似欠測データセットにおいて有効に作用しており、これは病院情報システムに蓄積された、欠測値の多い検査結果のデータセットに対しても有効であることが考えられる。また、本研究の様々な結果は、波形情報の前処理や機械学習アルゴリズムの選択などの際の一つの基準として寄与することが期待できる。

6. 結論

蛋白分画検査は、現在オーダー数が減少傾向にある検査であるが、非常に安価な検査であり、かつ得られる情報量は多

い。本研究にて、機械学習モデルと組み合わせることにより、特定の検査項目に対しては予測性能が高く、欠測値補間などの疫学研究用途における利用可能性を示すことができた。

参考文献

- 1) 藤田清貴. 第2章血清蛋白異常症の分析法. 臨床検査で遭遇する異常蛋白基礎から発見・解析法まで. 医歯薬出版, 2010 : 13-8
- 2) Weber GM, Kohane IS. Extracting physician group intelligence from electronic health records to support evidence based medicine. PLoS one 2013 ; 8(5) : e64933.
- 3) Zhang Z, Goyal H, Lange T, Hong Y. Healthcare processes of laboratory tests for the prediction of mortality in the intensive care unit: a retrospective study based on electronic healthcare records in the USA. BMJ open 2019 ; 9(6) : e028101.
- 4) Dohoo IR. Dealing with deficient and missing data. Preventive veterinary medicine 2015 ; 122(1-2) : 221-8.
- 5) Rubin, DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons 2014 ; 81

表2 各対象検査項目における、各機械学習モデルの AUC 値および精度。LR: ロジスティック回帰、L1LR: L1 正則化付きロジスティック回帰、SVM: サポートベクターマシン、CNN: 一次元畳み込みニューラルネットワーク、RF: ランダムフォレスト、GBDT: 勾配ブースティング決定木、AUC: area under the curve(AUC 値)、ACC: accuracy(精度)。灰色の網掛けは各検査項目において AUC が最高値のものを示す。*: LR と比較し、95%信頼区間が重複しない項目。

検査項目名称	LR		L1LR		SVM		CNN		RF		GBDT	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
アルブミン	0.929 [0.903–0.954]	0.86	0.938 [0.912–0.962]	0.87	0.975 [0.957–0.989]	0.98	0.978* [0.961–0.991]	0.93	0.962 [0.944–0.978]	0.89	0.970 [0.955–0.984]	0.91
白血球数	0.643 [0.580–0.705]	0.73	0.664 [0.602–0.723]	0.75	0.674 [0.613–0.735]	0.67	0.699 [0.642–0.757]	0.75	0.713 [0.655–0.767]	0.76	0.723 [0.667–0.783]	0.74
赤血球数	0.737 [0.690–0.781]	0.71	0.797 [0.754–0.834]	0.75	0.838 [0.799–0.874]	0.84	0.832 [0.792–0.870]	0.76	0.848 [0.812–0.885]	0.79	0.860* [0.823–0.895]	0.79
血小板数	0.602 [0.543–0.660]	0.66	0.638 [0.577–0.692]	0.68	0.782 [0.734–0.830]	0.78	0.759 [0.709–0.805]	0.75	0.755 [0.703–0.806]	0.75	0.783* [0.736–0.824]	0.77
ヘマトクリット	0.789 [0.746–0.828]	0.72	0.822 [0.781–0.862]	0.75	0.868 [0.832–0.901]	0.87	0.862 [0.826–0.896]	0.80	0.859 [0.824–0.893]	0.78	0.869* [0.835–0.900]	0.80
ヘモグロビン	0.774 [0.728–0.817]	0.72	0.824 [0.787–0.862]	0.77	0.851 [0.814–0.887]	0.85	0.842 [0.806–0.879]	0.78	0.862 [0.827–0.895]	0.79	0.876* [0.846–0.906]	0.80
クレアチニン	0.668 [0.620–0.720]	0.63	0.709 [0.662–0.759]	0.68	0.789* [0.745–0.830]	0.79	0.787 [0.741–0.830]	0.73	0.785 [0.742–0.829]	0.72	0.783 [0.738–0.825]	0.72
尿素窒素	0.735 [0.676–0.783]	0.74	0.758 [0.707–0.806]	0.77	0.797 [0.748–0.845]	0.80	0.817 [0.768–0.862]	0.80	0.802 [0.756–0.846]	0.79	0.823 [0.782–0.867]	0.81
尿酸	0.582 [0.515–0.648]	0.72	0.644 [0.580–0.711]	0.78	0.667 [0.609–0.724]	0.67	0.669 [0.602–0.732]	0.76	0.694 [0.632–0.758]	0.77	0.702 [0.638–0.763]	0.77
総ビリルビン	0.597 [0.520–0.685]	0.83	0.689 [0.616–0.754]	0.84	0.723 [0.657–0.782]	0.72	0.695 [0.614–0.764]	0.85	0.740 [0.670–0.803]	0.84	0.705 [0.642–0.771]	0.85
AST	0.604 [0.539–0.667]	0.72	0.659 [0.601–0.720]	0.77	0.718 [0.666–0.769]	0.72	0.711 [0.658–0.764]	0.75	0.700 [0.646–0.753]	0.76	0.701 [0.643–0.757]	0.74
ALT	0.612 [0.544–0.677]	0.72	0.682 [0.622–0.743]	0.76	0.752 [0.695–0.809]	0.75	0.674 [0.613–0.735]	0.74	0.752 [0.699–0.807]	0.76	0.773* [0.719–0.824]	0.79
アルカリホスファターゼ	0.676 [0.606–0.739]	0.78	0.743 [0.674–0.802]	0.82	0.793 [0.732–0.848]	0.79	0.762 [0.699–0.817]	0.84	0.760 [0.698–0.818]	0.83	0.763 [0.703–0.822]	0.84
γ-GT	0.650 [0.590–0.711]	0.72	0.664 [0.605–0.722]	0.72	0.715 [0.662–0.767]	0.72	0.672 [0.613–0.730]	0.71	0.725 [0.669–0.777]	0.73	0.749 [0.695–0.804]	0.77
ナトリウム	0.732 [0.672–0.79]	0.78	0.788 [0.730–0.847]	0.81	0.768 [0.708–0.823]	0.77	0.774 [0.708–0.829]	0.81	0.813 [0.758–0.863]	0.83	0.812 [0.756–0.866]	0.80
カリウム	0.721 [0.647–0.791]	0.83	0.723 [0.637–0.802]	0.88	0.786 [0.709–0.851]	0.79	0.705 [0.619–0.784]	0.89	0.782 [0.704–0.850]	0.89	0.782 [0.706–0.846]	0.89
C反応性蛋白	0.855 [0.826–0.880]	0.80	0.891 [0.858–0.918]	0.80	0.901 [0.869–0.929]	0.83	0.886 [0.852–0.918]	0.82	0.885 [0.851–0.915]	0.80	0.912* [0.883–0.938]	0.91
グルコース	0.68 [0.622–0.736]	0.67	0.716 [0.663–0.765]	0.70	0.758 [0.706–0.807]	0.76	0.744 [0.693–0.792]	0.72	0.721 [0.666–0.775]	0.72	0.763 [0.714–0.808]	0.73
アミラーゼ	0.582 [0.515–0.652]	0.72	0.578 [0.505–0.651]	0.74	0.632 [0.564–0.695]	0.63	0.656 [0.59–0.721]	0.73	0.704 [0.639–0.77]	0.77	0.684 [0.618–0.744]	0.76
クレアチンホスホキナーゼ	0.649 [0.591–0.710]	0.66	0.724 [0.671–0.777]	0.73	0.751 [0.700–0.801]	0.75	0.749 [0.696–0.799]	0.74	0.749 [0.699–0.797]	0.70	0.769* [0.719–0.815]	0.65
総コレステロール	0.68 [0.616–0.744]	0.72	0.745 [0.692–0.798]	0.75	0.827* [0.780–0.873]	0.83	0.802 [0.750–0.853]	0.80	0.799 [0.747–0.848]	0.78	0.800 [0.749–0.852]	0.78

表3 各対象検査項目における、各機械学習モデルの推定されたパラメータの一覧。SVM: サポートベクターマシン、CNN: 一次元畳み込みニューラルネットワーク、RF: ランダムフォレスト、GBDT: 勾配ブースティング決定木。ここに示していないパラメータは、ライブラリのデフォルト値を使用。

検査項目名称	最良学習モデル	パラメータ					
アルブミン	CNN	filters: 16	kernel size: 4	pool size: 2	dropout rate: 0.2	epoch: 500	-
白血球数	GBDT	learning rate: 0.001	lambda L1 : 0.183	lambda L2 : 0.002	num_leaves : 31	max depth : 10	feature fraction: 0.6
赤血球数	GBDT	learning rate: 0.01	lambda L1 : 0.003	lambda L2 : 0.004	num_leaves : 31	max depth : 10	feature fraction: 0.8
血小板数	GBDT	learning rate: 0.01	lambda L1 : 0.501	lambda L2 : 0.329	num_leaves : 31	max depth : 9	feature fraction: 0.3
ヘマトクリット	GBDT	learning rate: 0.01	lambda L1 : 0.001	lambda L2 : 2.014	num_leaves : 15	max depth : 7	feature fraction: 1
ヘモグロビン	GBDT	learning rate: 0.01	lambda L1 : 0.005	lambda L2 : 0.011	num_leaves : 15	max depth : 10	feature fraction: 0.2
クレアチニン	SVM	Cost: 2.572	γ : 0.006	-	-	-	-
尿素窒素	GBDT	learning rate: 0.01	lambda L1 : 0.002	lambda L2 : 0.016	num_leaves : 31	max depth : 8	feature fraction: 0.6
尿酸	GBDT	learning rate: 0.005	lambda L1 : 0.109	lambda L2 : 0.012	num_leaves : 31	max depth : 8	feature fraction: 0.3
総ビリルビン	RF	min split: 10	max nodes: 48	-	-	-	-
AST	SVM	Cost: 1.224	γ : 0.009	-	-	-	-
ALT	GBDT	learning rate: 0.01	lambda L1 : 0.046	lambda L2 : 0.035	num_leaves : 15	max depth : 9	feature fraction: 0.2
アルカリホス ファターゼ	SVM	Cost: 1.387	γ : 0.009	-	-	-	-
γ -GT	GBDT	learning rate: 0.005	lambda L1 : 0.123	lambda L2 : 0.367	num_leaves : 31	max depth : 10	feature fraction: 0.3
ナトリウム	RF	min split: 13	max nodes: 64	-	-	-	-
カリウム	SVM	Cost: 23.29	γ : 0.046	-	-	-	-
C反応性蛋白	GBDT	learning rate: 0.01	lambda L1 : 3.335	lambda L2 : 6.439	num_leaves : 7	max depth : 7	feature fraction: 0.9
グルコース	GBDT	learning rate: 0.005	lambda L1 : 0.331	lambda L2 : 0.013	num_leaves : 31	max depth : 8	feature fraction: 0.4
アミラーゼ	RF	min split: 14	max nodes: 56	-	-	-	-
クレアチンホスホ キナーゼ	GBDT	learning rate: 0.001	lambda L1 : 0.005	lambda L2 : 0.03	num_leaves : 15	max depth : 5	feature fraction: 0.5
総コレステロール	SVM	Cost: 3.107	γ : 0.007	-	-	-	-