機械学習を用いた検査結果からの患者取り違え採血検出手法の検討

Detecting patient mix-up on blood samples with machine learning

三谷 知広[*1] Mitani Tomohiro　土井 俊祐[*2] Doi Shunsuke　横田 慎一郎[*2] Yokota Shinichiroh　今井 健[*1] Imai Takeshi　大江 和彦[*1*2] Ohe Kazuhiko

[*1] 東京大学大学院医学系研究科医療情報学分野
Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo

[*2] 東京大学医学部附属病院企画情報運営部
Department of Healthcare Information Management, The University of Tokyo Hospital

Patient mix-up on blood samples is one of the common causes of blood test errors. It is also known as patient misidentification problem. Although the detection of mix-up is commonly performed by naive comparison with the last laboratory results of the same patients: delta checks, either the sensitivity or the specificity of delta checks is not satisfactory. To establish a new detection system, we made simulated mix-up data from actual data of complete blood count and serum chemistry in our hospital. Using differences from the previous laboratory results as features, a highly accurate detection system was built by machine learning technique. An XGBoost model recorded the best ROC AUC score of 0.9986.

1. はじめに

誤った患者からの採血は再検査や誤診などの原因となり，提出された検体の 0.1%〜0.3%に発生すると言われている[Dunn 2010, Schifman 2017]. このような患者取り違えはバーコードによるラベルや採血時の患者認証，複数の認証方法の導入などにより減少するが，人員不足や疲労，作業の中断などに起因するプロトコル違反によって発生するとされる[Randell 2019].

検体検査結果のチェック機構としては、各項目が前回値とどの程度変化したかを比較するデルタチェックと呼ばれる手法が広く用いられているが，その精度は十分とは言えない. デルタチェックにおける検出アラートの大半は生理的な変化や治療に関連する変化であり，検査エラーの関連する検出アラートは全体の約 4.8%である. さらにその大半は試薬の干渉やサンプルの汚染によって引き起こされ，患者取り違えに起因する検出アラートはアラート全体の約 0.3%であったと報告されている[Schifman 2017]. また，最近のレビューにおいて，多くの一般的なデルタチェックで感度は 30%未満であったと報告されている[Rendall 2019]. 単純なデルタチェックに代わる方法として，どの項目がデルタチェック陽性になったかに応じて検体としての判定を変えるルールベースの手法[Miller 2015]や，検査値ごとにデルタチェックの閾値を変える手法[Sourati 2015]，後述する加重累積デルタチェック法[Yamashita 2013]などが提案されているが，上記の中で最も精度の良い Yamashita らの報告でも ROC 曲線下面積(ROC AUC)は 0.98 であり，真のエラー率を 0.1%と仮定すると感度 0.90 を得る場合の陽性適中率は約 1〜2%となる. 渉猟した範囲で，採血取り違え検体の検出に昨今の機械学習技術を活用した報告はなかった.

検査結果を用いた患者取り違え採血の検出は，異常検知および変化検知のタスクとして捉えることができるが、個人個人の採血結果の推移は多系列かつ多変量な時系列データであり、各系列の時間軸が一般的な時系列データに比べて短いという点で時系列データに対する異常検知手法を適応することは難しい. 本研究では人工的な取り違えデータを作成した上で，主に以前の値との差分を特徴量として使用し，古典的な異常検知手法である Mahalanobis-Taguchi 法および複数の機械学習技術を適用し，患者取り違え採血の検出を行なった.

連絡先：三谷知広，東京大学大学院医学系研究科医療情報学分野，東京都文京区本郷 7-3-1，tmitani-tky@umin.ac.jp

2. 対象データ

2.1 データの概要

東京大学医学部附属病院における 2011 年から 2017 年の血液検体による臨床検査を対象とした. データは患者の個人情報が匿名化された SS-MIX2 データベースより取得した. 生化学検査と血算検査など複数の検体採取容器にわかれていても一度の採血で実施されることが多く，今回は同一時刻に記録されたデータは一連の検査として扱った. なお，このデータは検査部門や担当医などにより不整合と判断されたデータは削除もしくは修正された後の最終的なデータである.

血液に対する全 374 万件の検査のうち，血算検査 8 項目(WBC, Hb, Hct, RBC, MCV, MCH, MCHC, Plt)および生化学検査主要 7 項目(Alb, AST, ALT, BUN, Cre, Na, K)を含む 258 万件の検査を抽出した. この 7 項目は，当院での各項目の検査数を元にシステムがカバーするデータ割合を元に決定した. なお，対象とならなかった 116 万件の大半は血液ガス検査および腫瘍マーカー検査や抗体検査，ホルモン検査，血液型検査などであった.

血算および生化学主要 7 項目を含む 258 万件の中で，欠損値が 4 項目以内の検査は 233 万件であった. 各患者の初回検査 43 万件を除いて，前回値を有する 190 万件を対象とした.

2.2 取り違えのシミュレーション

まず患者ごとに検査結果を時系列に並べて window size を 4 回とした sliding window を適用し，今回値，前回値〜前々々回値の計 4 回分の検査結果を持つ部分時系列データを作成した. このデータを元に検査時刻や以前の検査結果などはそのまま固定して今回の検査結果のみを後述の基準による他人の検査

結果に取り替えた擬似的な患者取り違え採血データを作成した. ここに元の取り替えていないデータを加え, 4 回分の部分時系列データから今回値が取り替えられたデータか本来のデータかを二値分類するタスクとした.

なお, 入院・外来の違いや病棟の違いにより疾患・病態の分布が異なり, 検査結果の分布も異なる. 全患者内で検査結果を取り替えると, そのような入院外来・病棟ごとの分布の違いを学習してしまい, 実運用時に機能しない懸念がある. 実際の患者取り違えの状況に近づけるべく, 同一部署内での採血結果を取り替えることとし, また, 欠損パターンの違いを学習することも避けるため, 部署に加えて 15 項目中の欠損パターンも同一なデータ間で取り替えることとした. 実装の問題から各グループ内での単純な順列並び替えとしたが, 同一患者のデータに交換されてしまったデータは5186件(0.27%)と少なく, これらは削除した.

また, train/dev/test set への分割については, train set: 2011 年～2017 年 7 月, dev set: 2017 年 8 月～10 月, test set: 2017 年 11 月～12 月として test set に出現した患者のデータは dev set および train set から取り除き, さらに dev set に出現した患者のデータは train data から取り除いた. 以上により train set 380 万件, dev set 4 万件, test set 9 万件のデータを作成した.

3. 方法

3.1 検討したモデル

(1) デルタチェック (baseline)

閾値としては当院内で使用している以下の閾値を用いた(表 1.). 当院におけるデルタチェックは患者取り違えのみではなく試薬のエラーなどを含めた検査エラーの検出を目的として用いられており, 今回対象とした項目以外にも設定項目がある. また, 全ての検査項目を網羅しているわけではない. 本研究で対象とした 15 項目中の 9 項目に基準が設定されており, 9 項目について基準外となった項目数を取り違え陽性・陰性判別の閾値としたときの感度・特異度をプロットして ROC AUC を計算した. また, 1 項目でも基準外のものを陽性と判定した場合の Accuracy を求めた.

表 1. 当院におけるデルタチェック基準(該当項目)

項目	下限	上限	項目	下限	上限
Plt	70%	200%	BUN	70%	150%
MCV	95%	105%	Cre	80%	120%
Hct	85%	115%	Na	96%	104%
ALT	20%	350%	K	81%	123%
AST	20%	350%			

(2) Weighted Cumulative Delta-Check Index (wCDI)

Yamashita らによって報告された患者取り違え検体検出手法である[Yamashita 2013]. 各検査項目の分布を Box-Cox べき乗変換によって標準正規分布に変換して前回値との差分をとると, 個人間変動に比して個人内変動が大きい項目は広い分布となり, 逆に個人内変動が小さい項目は狭い分布となる. よって検査に含まれる各項目について, 正規化した前回値との差分をその分散の逆数で重み付けした加重平均を異常度の指標とし, 患者取り違え採血検知に用いた. データの存在する項目についてのみ平均を取るため欠損値に対応可能なモデルであり, 論文中では計算に使用できた項目数ごとの精度が報告されていた. 10項目以上でほぼプラトーに達したとの報告であり, 今研究では 15 項目中 10 項目以上で差分の計算できるデータを対象とした. 実装には scikit-learn [Pedregosa 2011]を使用した.

(3) Mahalanobis-Taguchi 法

対数正規分布が既知の項目については対数変換をおこなった上で, 前回値との差分を計算した. MCV, MCHC, MCH は Hb, RBC, Hct から計算される値であり, 多重共線性の懸念からこの 6 項目中 Hb, MCH, MCV を用いた. 以上により得られた 12 項目の差分データから Mahalanobis-Taguchi 法による異常度を計算し、その異常度についての ROC AUC を計算した. Accuracy を計算した閾値は, dev set 上の accuracy が最良となる点に設定した. なお, このモデルはデータの欠損に対応していないことから, 今回値および前回値が完全データであるデータ(train 97 万件, dev 2.3 万件)のみを対象とした.

(4) Deep Neural Network

欠損なしのデータのみを使用して sliding window を作成しなおし, 4 回×15 項目すべてが揃ったデータ(train 104 万件, dev 2 万件)を対象とした. 一部の項目を対数化した上で全項目について平均 0 標準偏差 1 に正規化し, MLP, 1D CNN, LSTM を検討した. CNN は時系列分類タスクに対して精度が優れると報告される ResNet 構造をベースとした[Hassan 2018]. いずれも Tensorflow を使用し, 層数や中間層のサイズなどのハイパーパラメータは Optuna により探索した.

(5) Gradient Boosting Decision Trees (GBDT)

GBDT のアルゴリズムとしては広く使用されている XGBoost[Chen 2016]を用いた. 今回値と, 各前回値・前々回値・前々々回値との差・時間差, 今回の測定時刻を特徴量として用いた. なお, 前々回値は約 14.3%, 前々々回値は 25.5%で欠損していた.

3.2 モデルの評価方法

各ハイパーパラメータチューニングは dev set 上で行った. 各モデルは dev set の ROC AUC で比較し, 最良のモデルを最終モデルとした. 最終モデルに対して test set における精度評価を行った. また, 一例として, baseline であるデルタチェックと同じ感度を実現できる閾値を設定し, 混同行列を求めた.

3.3 軽量モデルの検討

XGBoost を用いたモデルに対して以下の検討を追加した.

(1) 特徴量の削減

特徴量を削減したデータセットでの学習結果を比較した.

(2) データ量の削減

検査実施件数が小規模な医療機関でも自施設のデータのみでモデルを構築できるか検討するため, 100 万件, 10 万件, 1 万件, 1000 件のデータについて学習させた結果を比較した.

4. 結果

4.1 モデルの比較

各モデルにおける dev set 上の ROC AUC を比較検討した. XGBoost は欠損値を含むデータを対象としていながら最良の結果を達成しており, 最終モデルとして XGBoost によるモデルを採用した. なお, デルタチェック (baseline)において 1 項目でも基準外の項目があった検査を取り違え陽性とした場合の感度は 95.9%, 特異度は 72.3%であった. 真の取り違え発生率を 0.1%と仮定した場合の陽性適中率は 0.35%であった.

表 2. 各モデルによる結果

方法	Accuracy	ROC AUC	備考
Delta Check	0.8400	0.9306	Baseline
wCDI	0.9248	0.9781	欠損値含む
MT	0.9304	0.9787	
MLP	0.9678	0.9958	
CNN	0.9686	0.9950	
LSTM	0.9705	0.9962	
XGBoost	**0.9792**	**0.9980**	欠損値含む

4.2 最終モデルの検出精度

XGBoost による最終モデルとデルタチェック(baseline)における test set 上の結果を示す(図 1). なお, 前項の結果は dev set 上の結果であり以下の test set における最終評価結果とは異なっている. Test set のデータはこの最終評価にのみ用いた.

XGBoost のモデルにおいて, 閾値を 0.05 刻みとして感度を計算し, デルタチェック(baseline)による感度 95.9%を達成できる閾値の下限である p=0.70 での混同行列を求めた(表 3). この設定において真の取り違え発生率を 0.1%と仮定した場合の陽性適中率は 11%であった.

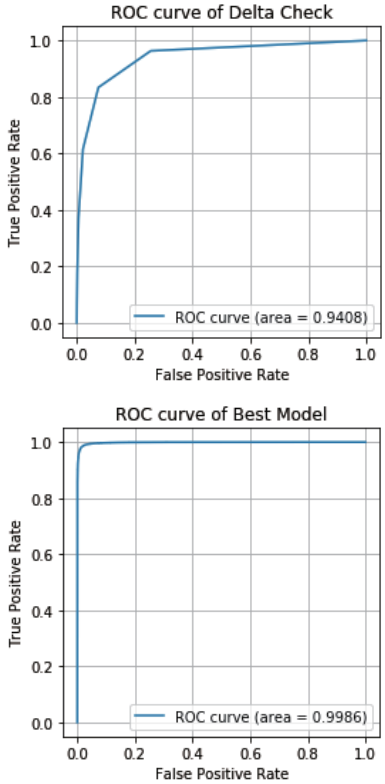

図 1. Test set における ROC 曲線
(上:baseline, 下:XGBoost を用いた最終モデル)

表 3. 最終モデルにおける 閾値を p=0.70 とした混同行列

混同行列		予測	
		取り違えなし	取り違えあり
実際	取り違えなし	48575 (99.3%)	367 (0.7%)
	取り違えあり	1526 (3.1%)	46913 (96.9%)

最終モデルにおける特徴量重要度上位 15 項目を図 2 に示した. 特徴量の重要度を計算する方法は複数知られているが, ここでは各ノードの損失減少量をノードごとに平均した値を出力した.

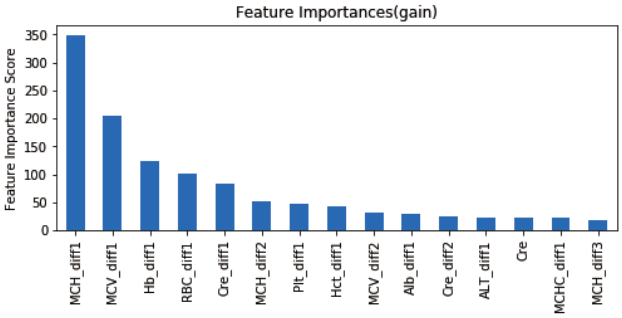

図 2. 最終モデルにおける特徴量重要度
(Suffix のついていない項目は今回値そのものを示し, suffix 末尾の数字は何回前との差分かを示す)

4.3 軽量モデルの検討

以下の実験において、learning_rate=0.3, n_estimators=200, max_depth=8 の設定で学習した.

(1) 特徴量の削減

表 4. 特徴量を減らした設定での ROC AUC

特徴量セット	ROC AUC
4 回分 (今回値+今回値との差)＋時刻・時間差	**0.9972**
4 回分 (今回値+今回値との差)	0.9959
2 回分 (今回値+今回値との差)＋時刻・時間差	0.9965
2 回分 (今回値+今回値との差)	0.9952
2 回分 (今回値+前回値そのまま)	0.9942
2 回分 (今回値+今回値との差, 血算のみ)	0.9800
2 回分 (今回値+今回値との差, 生化学のみ)	0.9701

(2) データ量の削減

表 5. 学習データ量を減らした設定での ROC AUC

学習データ量	ROC AUC
380 万件	**0.9972**
100 万件	0.9971
10 万件	0.9953
1 万件	0.9903
1000 件	0.9804

5. 考察

XGBoost は欠損値に対応したモデルであり, 適用可能なデータ数を維持しながら高精度の判別モデルを構築することができた. デルタチェック(baseline)の ROC AUC 0.9306, 既存研究である wCDI 法の ROC AUC 0.9781 に対し, ROC AUC 0.9980 を達成した. なお, DNN についてはいずれも正規化した 4 回分のデータをそのまま入力としており, 以前の値との差分をとって使用した XGBoost のモデルと単純に比較することはできない. 特徴量設計やネットワーク構造の工夫などで精度が改善する可能性はあるが, 欠損値への本質的な解決が難しいことから, さらなる検討は行わなかった.

wCDI は報告とほぼ同等の ROC AUC が再現された. ROC AUC は XGBoost のモデルより低かったが, 連続量で結果が得られる検査項目に対して汎用的に適用できる手法であり, 今研究のモデルでは対象としなかった腫瘍マーカー検査や抗体検

査，ホルモン検査などにも適用できるメリットがある．実運用では両者を組み合わせて使用することも検討に値する．

なお検体取り違えは稀な事象であり，一般に陽性的中率は低い．1 日に 1000 件検査があり 0.1%の確率で取り違えを生じるという状況を仮定すると，デルタチェックによる baseline では毎日 277 件のアラートのうちの 0.96 件が実際の取り違えであったという状況に相当し，陽性適中率は 0.35%にとどまる．閾値を 2 項目にあげると陽性適中率は 1.1%となるが感度は 83.7%に低下する．低い陽性的中率は不要なカルテレビューや再度の採血検査などを招くうえ，またここから取り違え検査を特定することに困難を伴うことが予想される．XGBoost による最終モデルで閾値を baseline と同等の感度 96%を維持できる p=0.70 に設定した場合，同様の条件として毎日 8.5 件の検出アラートのうち 0.97 件が実際の取り違えという状況に相当し，感度は 97%で陽性適中率は 11%である．なお，閾値を p=0.95 とすると感度は 92%に低下するものの，陽性適中率は 36%となる．これらはデルタチェックにより生じていた人手での確認による労力の削減や，再検査の減少などによる患者負担の減少に繋がりうる結果である．

特徴量としては，MCV や MCHC，Cre といった個人内変動の少ない項目の差分が重要であった．Cre のみ今回値そのものが重要な特徴量として TOP15 に入ったが，これは Cre の変化において差よりも比が意味を持つためと思われ，前回値との差分ではなく比を使用した方がよかったかもしれない．検査時刻や前回との時間差は，シミュレーションデータにおける人工的な取り違えの有無に独立な特徴量であり，これらの特徴量追加によって ROC AUC 値が向上したことは，XGBoost によって他の特徴量との交互作用がモデルに組み込まれたことを示唆する．血算のみ・生化学のみとスピッツごとのデータに分けると ROC AUC は低下し，特に生化学のデータのみとすると大きく低下した．実用的な精度を得るには血算と生化学を組み合わせるか，特に欠損値の多い生化学検査においては例えばいくつかの特定の項目は全例でルーチンに測定するなどの何らかの工夫が必要と思われる．

データ量については，1 万件～10 万件程度でも 0.99 以上の ROC AUC を達成できており，一般の病院が自施設で収集したデータからも実用的なモデルを構築可能と考えられる．なお，実際のデータには同一患者からのデータが多く含まれるが，今回の実験設定では train set 380 万件からランダムに抽出しており，データ量が等しくても異なった患者からのデータがより多く含まれる．データの多様性が増すため，実データから同等の精度を得るためにはより多くのデータを必要とするかもしれない．

本研究の limitation としては，SS-MIX2 標準化ストレージ内のデータのみを用いたため輸血や手術といった臨床情報を組み込めなかった点が挙げられる．ROC AUC はデルタチェックや wCDI 法に比べ改善したものの，陽性適中率は前述の設定で 11%である．柔軟な特徴量設計が可能という GBDT のメリットを活かして直前の輸血や手術の有無，緊急検査かなどの特徴量の追加を試み，陽性適中率の向上を図りたい．また，実運用に向けた課題として個々の結果に対する判断根拠の可視化などを検討している．

6. まとめ

今回の研究では，機械学習による患者取り違え採血の検出を試みた．これまでは前回値との単純な比較であるデルタチェックが推奨されていたが，機械学習の手法を用いることで感度・特異度ともに大幅に向上した．

参考文献

[Chen 2016] Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754, 2016.

[Dunn 2010] Dunn, E. J. and Moga, P. J.:Patient misidentification in laboratory medicine: a qualitative analysis of 227 root cause analysis reports in the Veterans Health Administration, Archives of Pathology & Laboratory Medicine, 2010.

[Fawaz 2018] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L. and Muller, P. A.: Deep Learning for time series classification: a review. arXiv:1809.04356, 2018.

[Miller 2015] Miller, I.: Development and Evaluation of a Logical Delta Check for Identifying Erroneous Blood Count Results in a Tertiary Care Hospital., Archives of Pathology & Laboratory Medicine, 2015.

[Pedregosa 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., … Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011.

[Rendall 2019] Rendall, E. W. and Yenice, S.: Delta Checks in the clinical laboratory (Review), Critical Reviews in Clinical Laboratory Sciences, 2019.

[Schifman 2017] Schifman, R. B., Talbert, M. and Souers, R. J.: Delta Check Practices and Outcomes: A Q-Probes Study Involving 49 Health Care Facilities and 6541 Delta Check Alerts, Archives of Pathology & Laboratory Medicine, 2017.

[Sourati 2015] Sourati, J., Erdogmus, D., Akcakaya, M., Kazmierczak, S. C. and Leen, T. K.: A Novel Delta Check Method for Detecting Laboratory Errors. IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015.

[Yamashita 2013] Yamashita, T., Ichihara, K. and Miyamoto, A.: A Novel Weighted Cumulative Delta-Check Method for Highly Sensitive Detection of Specimen Mix-up in the Clinical Laboratory. Clinical Chemistry and Laboratory Medicine, 2013.