

A New Ensemble Learning Algorithm for ICU mortality prediction

LI NANFANG

集中治療室 (ICU) は、重症患者を監視し蘇生させるための特別なユニットです。複雑な病態、不安定なバイタルサイン、死亡リスクの高い患者さんは、入院の初期段階で目標とする治療・ケアプランを立てて、ICU に入室する必要があります。そのため、医師は患者の重症度や院内死亡のリスクを迅速かつ正確に判断し、医療計画に反映させる必要があるのです。このニーズに応えるため、研究者は、ICU での死亡リスクを評価するために、Acute Physiology and Chronic Health Score や Simplified Acute Physiology Score など、さまざまなスコアリングシステムを提案しています。これらのスコアリングシステムは、何度も更新・改良され、現在では ICU 病棟で広く使用されている。

臨床現場で広く使われているスコアリングシステムは、シンプルで使いやすい反面、自動化レベルが低く、統合が困難で、欠損データにも対応できない。

臨床の現場で使われているスコアリングシステムは、シンプルで使いやすいのですが、自動化が進んでおらず、統合が難しく、欠損値を扱えないという欠点があります。近年、電子カルテシステムの向上に伴い、ICU 患者の死亡予測タスクに機械学習技術が適用され、臨床的な死亡リスクスコアリングモデルよりも優れた性能が示されています。しかし、ICU 患者の生理学的パラメータの複雑さと数の多さ、欠損データの多さ、死亡サンプルと生存サンプルの比率のアンバランスさなどが、予測モデル構築の難しさとして残っています。本論文では、これらの問題に対する解決策を提案する。

1. 本文のデータは、MIMIC-IV (Medical Information Mart for Intensive Care-IV v2.0) データベースから取得したものである。本論文では、臨床採点システム、クリティカルケア文献、医師の経験を組み合わせ、ICU での臨床判断によく使われる 23 の特徴を選定した。選択された特徴の妥当性を検証するために、特徴セットに含まれる各特徴のモデルへの寄与度を分析した。さらに、3 つの臨床モデルで使用されている急性期生理学スコアと慢性期健康スコアを含む生理学的パラメータを参照に、比較のための 3 つの特徴セットを構築した。実験の結果、選択した特徴で学習したモデルの方が性能が良く、医師の経験とも一致することが示された。

2. 欠損データの問題に対して、本文では、深層埋め込みクラスタリングに基づく K-最近傍補間法を提案する。このアルゴリズムは、深層埋め込みクラスタリングを中核として、多重クラスタリングによりサンプル最近傍行列を構築し、欠損サンプルの最近傍サンプルを K 個選択し、これらの最近傍サンプルの固有値の平均値で欠損を埋めるものです。本文では、データ補間法の検証のために、5260 個の非欠測標本を用い、非欠測標本に欠損値を追加し、非欠測標本と欠損標本との余弦類似度を比較することによって欠損標本を構成した。その結果、平均補間法、中央値補間法と比較して、補間されたデータは元のデータとの類似度が最も高いことがわかった。

3. 偏ったデータ分類の問題を解決するために、本論文ではクラスタリングに基づく適応的リサンプリング統合学習モデル (CAREnsemble) を提案する。このアルゴリズムは、まず学習セットの多数クラスのサンプルをクラスタリングし、各クラス内のサンプル数に応じて再サンプリングする。次に、再サンプリングしたサンプルと少数クラスのサンプルを組み合わせ、分布のバランスがとれた複数の学習サブセットを作り、これらのサブセットを用いて複数の弱い分類器を学習し、最後に Easy Ensemble 法によりそれらを統合する。7696 個のテストサンプルに対する本アルゴリズムの被験者作業特性曲線下面積、再現率、精度はそれぞれ 0.8684、80.86%、74.51%であり、RUSBoost、Adaboost、サポートベクターマシンといったモデルを超える性能であった。(行動統計科学)

キーワード: 死亡率予測, Deep Embedded Clustering, KNN, Ensemble learning