

厚生労働科学研究費補助金
(政策科学総合研究事業(統計情報総合研究事業))

令和7年度 分担研究報告書

ICD-11 の適用を通じて我が国の死因・疾病統計の向上を目指すための研究
「主疾患推定モデルの開発と特徴量設計—DPC/PDPS レコードを教師データとして—」

研究分担者 野口晴子 早稲田大学政治経済学術院 教授

研究協力者 青木志穂 早稲田大学理工学術院 博士後期課程 2 年

研究協力者 木内絢美 早稲田大学理工学術院 修士課程 2 年

研究要旨

レセプト情報・特定健診等情報データベース(National Database:NDB)は、多様な疾患に関する医療行為や医療費情報を網羅的に分析できる潜在力を有している。一方で、レセプト病名が患者の「主疾患」を必ずしも正確に反映していない可能性が指摘されており、疾患別分析を行う上での課題となっている。

そこで、本研究では、包括医療費支払制度(DPC/PDPS)において主傷病情報が付与されているレコードを教師データとして活用し、医科レセプト情報から主疾患を推定する機械学習手法の開発を行った。

本年度は、主として、特徴量設計、不均衡データへの対応手法、および解析基盤整備に関する検討を実施した。具体的には、レセプト電算コードを YJ コードへ変換し、薬効分類コードを特徴量として利用する前処理手法を構築した。また、不均衡データへの対応として、階層型二値分類および二段階分類のアルゴリズム構造を実装した。さらに、薬効分類コードの階層構造を活用した特徴量設計や、投与経路情報を特徴量として利用する方法についても検討を行った。

加えて、大規模レセプトデータ解析を想定し、PostgreSQL を用いたデータベース設計、Python によるデータ処理パイプライン、GitHub を用いたコード管理環境等の解析基盤整備を進めた。さらに、将来的な発展可能性として、Graph Convolutional Network(GCN)等のグラフベース機械学習手法についても技術的検討を行った。

以上より、本年度は、主疾患推定モデルの構築に向けた前処理手法、特徴量設計、不均衡データへの対応方法、分類アルゴリズム構造および解析基盤整備について、機械学習モデル開発の基盤整備を実施した。なお、本研究では、NDB 用いた結果の創出までには至らなかった。

A. 研究目的

レセプト情報・特定健診等情報データベース(National Database:NDB)は、罹患者数の多い疾患から、患者数が極めて少ない難病に至るまで、治療行為のパターンや詳細な医療費を分析できる潜在力を秘めている。一方で、NDB に収載されているレセプト病名は、

患者の「主疾患」を必ずしも正確に反映していない可能性が指摘されており、疾患別分析を行う上での課題となっている。

そこで、本研究では、包括医療費支払制度(DPC/PDPS)において主傷病情報が付与されているレコードを教師データとして活用し、

医科レセプト情報から主疾患を推定する機械学習手法の開発を目的とする。

特に、本研究では、NDB に特有な課題である「不均衡データ」への対応や、レセプト電算コードを薬効分類コードへ変換する特徴量エンジニアリングに着目し、主疾患推定精度の向上を目指す。さらに、階層型二値分類や二段階分類など、複数の分類アルゴリズムを比較・検討することで、主疾患推定に適したモデル構造を検討する。

また、将来的には、薬効分類コードの階層的表現や投与経路情報等を追加した特徴量設計を行うことで、より高精度かつ解釈可能性の高い主疾患推定モデルの構築を目指す。(図 1 参照)。

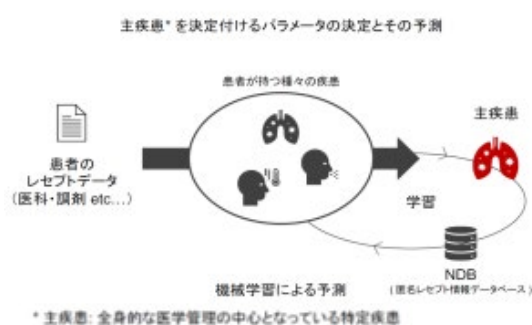


図 1 研究目的の概念図

B. 研究方法

本研究では、包括医療費支払制度(DPC/PDPS)において主傷病情報が付与されているレコードを教師データとして用い、医科レセプト情報から主疾患を推定する機械学習手法の開発を行った。

まず、レセプトデータに含まれる薬剤コードについて、レセプト電算コードから YJ コードへの変換を行い、さらに薬効分類コードを抽出した。レセプト電算コードは薬効そのものを反映したコード体系ではないため、同一薬効であっても製薬企業ごとに異なるコードが付

与される場合がある。そこで、本研究では、薬効分類番号を用いて特徴量を再構成し、one-hot encoding を行うことで、薬効に基づく特徴量エンジニアリングを実施した。

次に、NDB データが大きく不均衡なクラス分布を有する点に着目し、不均衡データへの対応手法として、階層型二値分類および二段階分類の検討を行った。

階層型二値分類では、データ数の多い疾患群から順に二値分類器を適用し、多段階的に疾患分類を行った。一方、二段階分類では、各疾患群について「当該疾患であるか否か」の二値分類を実施し、その予測結果または予測確率を新たな特徴量として用いた上で、最終的な多値分類を行った。

また、薬効分類コードについては、単純な one-hot encoding のみではなく、コード体系が持つ階層構造を活用する前処理手法についても検討した。具体的には、薬効分類コードを桁ごとに分割し、各階層情報を特徴量として利用する方法を検討した。

さらに、将来的な精度向上に向けて、内服薬・注射薬・外用薬等の投与経路情報を特徴量へ追加する手法についても検討を行った。

機械学習アルゴリズムとしては、XGBoost を中心に検討を行い、分類精度の比較評価を実施した。

また、本研究では、大規模レセプトデータを安全かつ効率的に解析するための計算環境およびデータ処理基盤の整備も実施した。具体的には、PostgreSQL を用いたデータベース設計、Python によるデータ処理パイプラインの構築、GitHub を用いたコード管理環境の整備等を行った。

さらに、大規模データ解析を念頭に、Docker コンテナや C++ベースの解析基盤を含む高速解析環境についても検討を行った。

C. 研究結果

本年度は、主疾患推定モデルの構築に向けた機械学習手法の開発および前処理手法の検討を中心に実施した。

まず、レセプト電算コードを YJ コードへ変換し、薬効分類コードを特徴量として利用する前処理手法を構築した。これにより、同一薬効を有する薬剤を統一的に扱うことが可能となり、薬効ベースでの特徴量設計手法を整理した。

次に、不均衡データへの対応として、階層型二値分類および二段階分類のアルゴリズム構造を実装した。階層型二値分類では、データ数の多いクラスから順次分類する構造を構築した。また、二段階分類では、各疾患群について実施した二値分類結果を特徴量として再利用する構造を実装した。

さらに、薬効分類コードの階層構造に着目し、コードの桁情報を段階的に利用する特徴量設計について検討を行った。加えて、内服薬・注射薬・外用薬等の投与経路情報を特徴量へ追加する方法についても検討を進めた。

また、本年度は、解析コードおよび関連資料の管理体制整備を行い、GitHub を用いたコード共有環境や、大規模データ解析を想定したデータベース・パイプライン構成についても整理を行った。さらに、Docker コンテナや C++ ベースの解析基盤を含む高速解析環境についても検討を実施した。

以上の結果、本年度は、主疾患推定に向けた前処理手法、特徴量設計、不均衡データへの対応方法、分類アルゴリズム構造および解析基盤整備について、機械学習モデル開発の基盤を整備した。なお、本研究では、NDB 用いた結果の創出までには至らなかった。

D. 考察／E. 結論

本研究では、医科レセプト情報から主疾患を推定する機械学習手法の開発を目的として、特徴量設計、不均衡データへの対応方法、および解析基盤整備を中心に検討を行った。

レセプトデータは、本来、診療報酬請求を目的として作成される情報であり、疾患分類を目的として構造化されたデータではない。そのため、主疾患推定を行う上では、薬剤コード、診療行為、副傷病等をどのように特徴量として表現するかが重要となる。本研究では、レセプト電算コードを薬効分類コードへ変換する特徴量設計を導入し、薬効ベースで情報を整理する前処理手法を検討した。

また、NDB データはクラス分布の偏りが大きく、不均衡データへの対応が重要な課題である。本研究では、階層型二値分類および二段階分類を導入することで、不均衡データに対応した分類構造を検討した。さらに、薬効分類コードの階層構造や投与経路情報を特徴量へ組み込む方法についても検討を行った。

加えて、本研究では、大規模レセプトデータ解析を念頭に、データベース設計、データ処理パイプライン、コード共有環境等の解析基盤整備も実施した。これらの基盤整備は、今後、大規模データを継続的かつ再現性を担保しながら解析する上で重要であると考えられる。

また、将来的な発展可能性として、Graph Convolutional Network (GCN) 等のグラフベース機械学習手法についても技術的検討を行った。レセプトデータは、疾患、薬剤、診療行為等の複雑な関連構造を有しており、今後はグラフ構造を活用したモデル開発も有望であると考えられる。

以上より, 本年度は, 主疾患推定モデル構築に向けた機械学習手法および解析基盤整備に関する基礎的検討を実施した. 今後は, 特徴量設計および分類モデルの改良を進めるとともに, モデル性能評価および実装可能性について更なる検討を行う必要がある.

F. 健康危険情報

特に無し.

G. 研究発表

1. 論文発表

特に無し.

2. 学会発表

特に無し.

H. 知的財産権の出願・登録状況(予定を含む)

1. 特許取得

特に無し.

2. 実用新案登録

特に無し.

3. その他

特に無し.

参考文献

Scikit-learn – Machine Learning in Python.

<https://scikit-learn.org/stable/>

(閲覧日:2024 年 5 月 24 日)

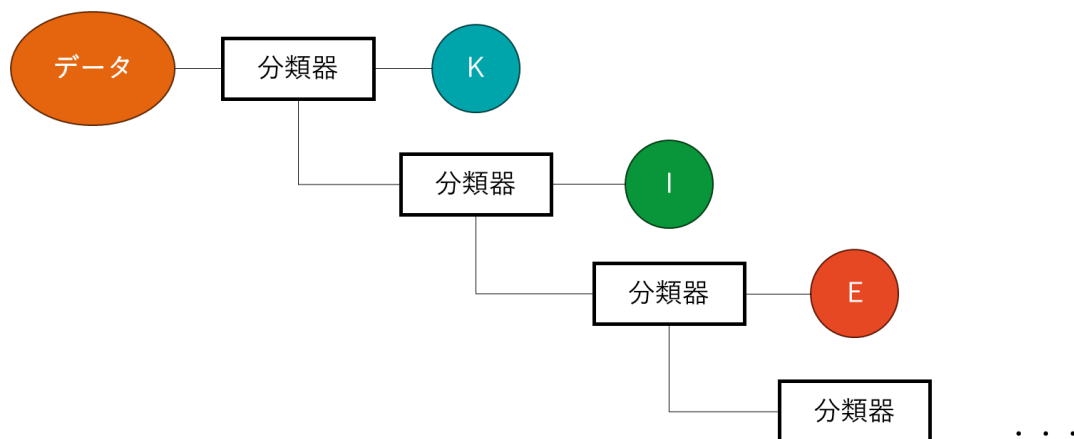


図2 階層型二値分類の構造

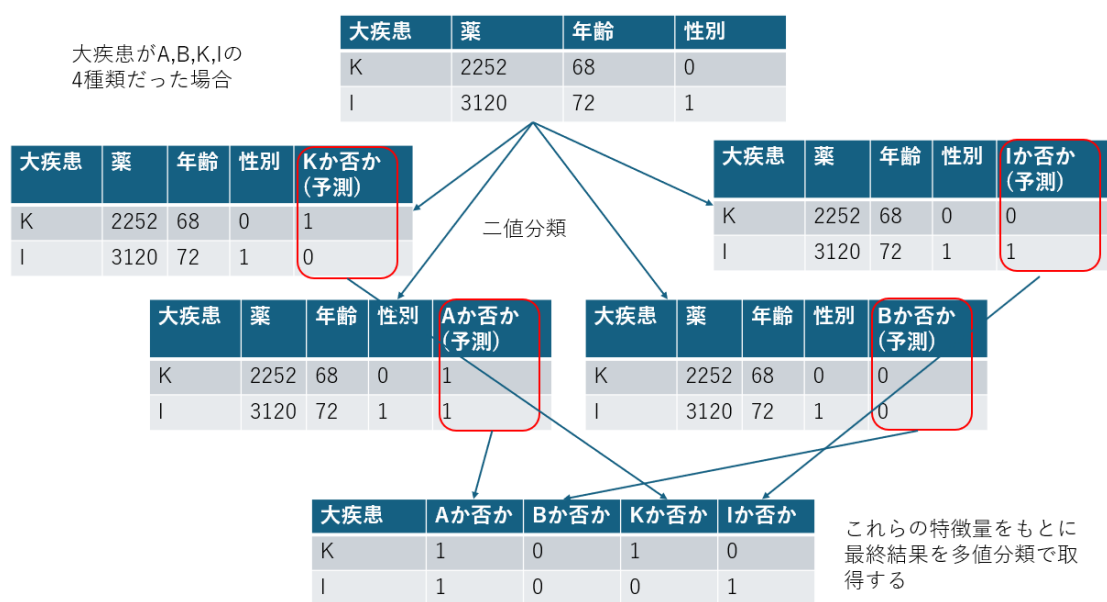


図3 二段階分類の例