Using Small Language Models to Reverse-Engineer Machine Learning Pipelines Structures

Nicolas Lacroix , Fr ´ ed ´ eric Precioso Universit ´ e Cˆote d'Azur, Inria, CNRS, I3S Sophia Antipolis, France firstname.name@univ-cotedazur.fr

Mireille Blay-Fornarino Universit ´ e Cˆote d'Azur, CNRS, I3S Sophia Antipolis, France firstname.name@univ-cotedazur.fr

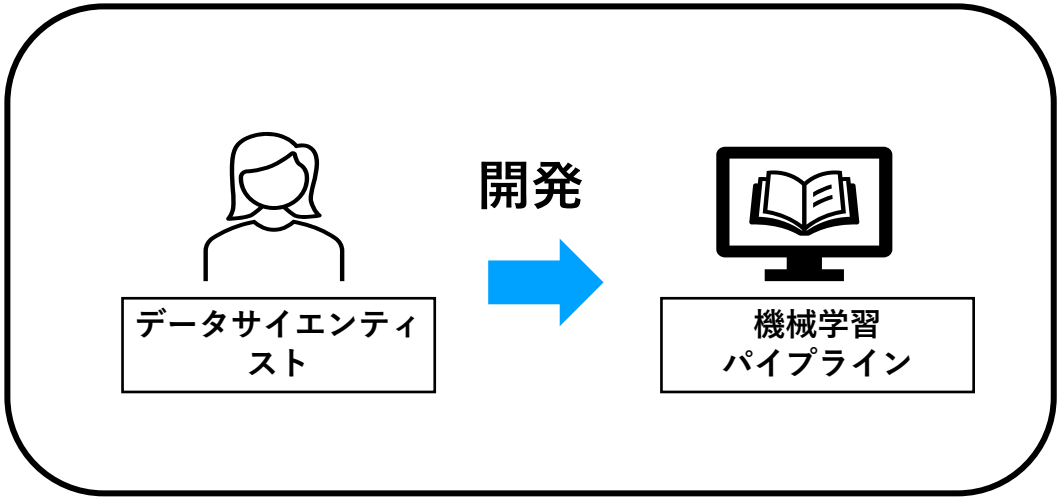
S ´ ebastien Mosser McSCert, McMaster University Hamilton, Ontario, Canada mossers@mcmaster.ca

SANAR 2026

学生番号：s2620113

発表者:福井　遥陽

背景



→どの順番に機械学習モデルを開発しているのか？

- パイプラインの品質向上や進化を支援するうえで重要
  - ワークフローがどう構成され，反復され，再利用されているか

背景

従来手法

- 特定ライブラリとパイプライン段階の対応表を用いた静的マッピング手法

自動化手法へと
遷移

- ルールベース手法
- 教師あり分類器
  - 特定ライブラリしか使えない（新しいライブラリに未対応）
  - 特定パイプライン構造しか対応できない
  - 性能が不十分

背景

従来手法

- 特定ライブラリとパイプライン段階の対応表を用いた静的マッピング手法

自動化手法へと
遷移

- ルールベース手法
- 教師あり分類器

**言語モデルを利用した
モデルに注目**

GPUの計算資源制限のためSLMに期待
→SLMを用いたMLパイプライン構造抽出は未検証

研究目的

・MLパイプラインの構造抽出に対するSLMの有効性を検証

- RQ1
  - SLMは既存分類器より高精度か
- RQ2
  - taxonomy（分類体系）の単語の選び方は分類結果に有意な影響を与えるか
  - 例）分類ラベルの表現（例：Training / Model Building）によって結果は変わるか
- RQ3
  - SLMを使うことで既存研究とは異なる実践パターンが見つかるか

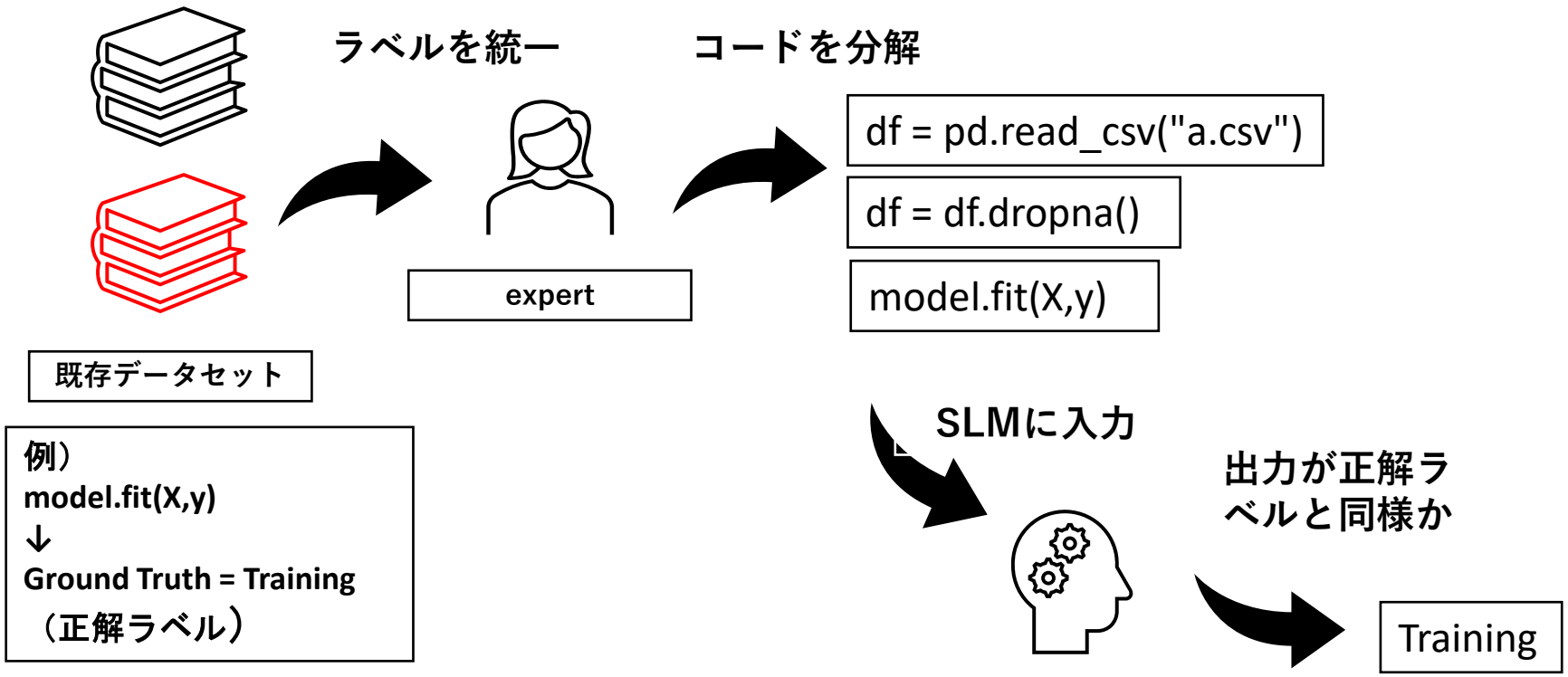
研究目的

・MLパイプラインの構造抽出に対するSLMの有効性を検証→帰無仮説で検証予定

- RQ1
  - SLMは既存分類器より高精度か
- RQ2
  - taxonomy（分類体系）の単語の選び方は分類結果に有意な影響を与えるか
  - 例）分類ラベルの表現（例：Training / Model Building）によって結果は変わるか
- RQ3
  - SLMを使うことで既存研究とは異なる実践パターンが見つかるか

実験設定

実験パイプライン



実験設定

評価指標

- Matthews Correlation Coefficient (MCC)
- F1-score
  - 分類器がどれだけ正確にpipline stageを予測できたか
- Cduration
  - 分類に要した時間
- SLMperplexity
  - モデルが予測にどれだけ不確実性を持っているか

実験設定

使用データセット

- Jupyter Notebook
  - 既存研究で利用されたラベル付きデータ

使用モデル

- 選定条件
  - オープンソースであること
  - モデルの重み尾が公開されていること
  - コード処理用に学習されたモデルであること
- 以上の条件で，humanevalベンチマークで最も性能の高い5つのモデル

実験設定

比較対象モデル

- 選定条件
  - コードからMLパイプライン構造を抽出していること
  - データサイエンティストの開発行動に関する分析結果を報告していること
  - 再現可能なオープンソース実装を公開していること

まとめ

本研究のまとめ

- 従来手法はルールベースや学習データに依存
- 本研究ではSLMを使用した構造抽出を検証
- 既存分類器との比較により有効性を検証

所見

- SANERで発表された実験計画主軸の論文
  - 実験設計の仕方がかなりしっかりしてる
- 有効性検証のために，影響要因を厳密に設定している
  - モデルのサイズ，プロンプト，推論設定，決定論的推論