大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

前処理パターンの抽出と評価

株式会社富士通研究所　西村 駿人　n.hayato@fujitsu.com

開発における問題点

- 機械学習応用システムの開発において，データの前処理は重要
- 前処理は専門的なスキルである
- 非専門家においても，データ前処理を容易に実施したい

手法・ツールの適用による解決

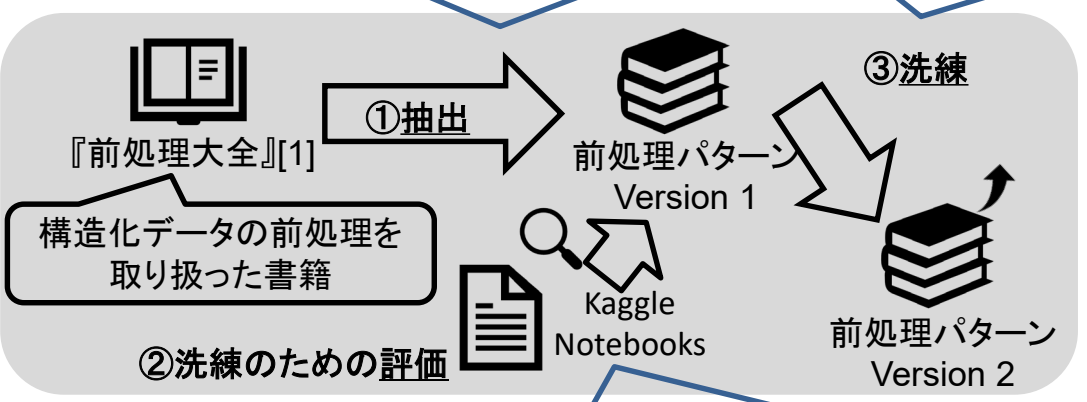
- IT技術者が利用可能な共通言語として**前処理パターン**を抽出
- 前処理パターン集の利用により前処理の容易な実施を目指す

前処理パターン抽出方法のアプローチ

発表者が3段階の工程で，前処理パターンの抽出と洗練を実施（①②③の順）

データ構造を対象とした前処理
（抽出，集約，結合，分割，生成，展開）を起点に抽出を実施

②評価 の結果を踏まえて
新たなパターン抽出や既存パターンの改版を実施



Vote数上位100件のうち，以下を満たす82件で人手で評価
- 構造化データ（CSVなど）を処理するもの
- Pythonソースコードが記載されているもの

③洗練で抽出したパターン

前処理パターン利用のためのパターン	外部情報を使ったデータ変換パターン
データの性質を記述情報で確認するパターン	データの性質をグラフで確認するパターン
特徴量に着目した表への変換パターン	評価のためのデータ分割パターン
ファイル形式変換パターン	データ拡張パターン
列の抽出パターン	データ縮小パターン
データの結合パターン	データ型変換パターン
行の抽出パターン	欠損値を補完するパターン
スケールを変更するパターン	数値の変換を行うパターン

[1]: 本橋 智光，“前処理大全”，技術評論社，2018

実施結果

①抽出 では11種類
③洗練 では16種類
のパターンを抽出

2パターンを除いて，パターンの出現を確認
（②のデータセットを利用）

Wikiシステムを利用し，
パターンをブラウザで編集・閲覧可能とした
→ユーザーはパターン集を参照し前処理実施が可能

評価

パターンの妥当性の確認のため，
各パターンの出現の有無を人手で確認

抽出に利用していないKaggle コンペティションのNotebookで出現確認の実施
- ■ Vote数上位のNotebookで頻出していた一部のパターンが非出現（データ確認関係）
- ■ Version2で新たに追加したパターンが出現
  →洗練活動が効果を発揮した

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

タイトル

所属　　名前　　メールアドレス（任意）

開発における問題点

ここでは，今回の修了制作で解決した開発における課題・問題点について述べる．例えば「XXシステム開発の際に，セキュリティ上の攻撃とそれに対する対策を系統的，網羅的に分析する必要がある」といったもの．「YYアプリに機能Aがなかった」等アプリケーション自体の問題点ではないことに注意する．

手法・ツールの適用による解決

ここでは，左で挙げた問題について，どのような手法・ツールの適用や提案によって解決したかを述べる．手法・ツールの名前を出すだけでなく，性質を明示し問題との対応がわかるようにする．例えば「達成目標を分析してシステムの構成要素を導出する系統的な方法を定めた要求分析手法KAOSを用いて・・・」といった感じ．

ポスターの構成

上の概要のタイトルも必要であれば変更してください．人によっては「適用」ではなく「拡張」「提案」「連携」等となると思います．ある程度であればマスタの方の大きさを変えていただいても構いません．

概要より下の部分の構成（スペースの分割等）についてはお任せします．フッタは残して下さい．

このフォーマットはA3になっていますが，実際にはポスターA0印刷，配布用A4印刷を行います．文字の大きさは最低13pt程度としてください．

注意点

Webにてアクセス制限なく公開するものであることに留意して下さい．

審査会での発表とは異なり，外部の方々・TopSEのツール・手法を知らない方々も対象となるため，
・モデリング方針
・ツールの設定
等の詳細よりも，
・その手法・ツールは何ができるか
・端的には，その図は何を表しており，その図を使って何をしたのか
といったことを概念的に説明することとなります．

提出

LMSから提出して下さい．

最終的には3月の修了式等のイベントにて，
（上司の方々や外部の来賓にも）掲示します．
・こちらで印刷時の様子を見て多少レイアウト等調整を行う可能性があります．
・印刷はNIIで（事務局が）行います．