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

AIを活かすためのデータ前処理方法の検討

株式会社NTTデータ アイ　勝木　啓介
NECソリューションイノベータ株式会社　佐々木　良

開発における問題点

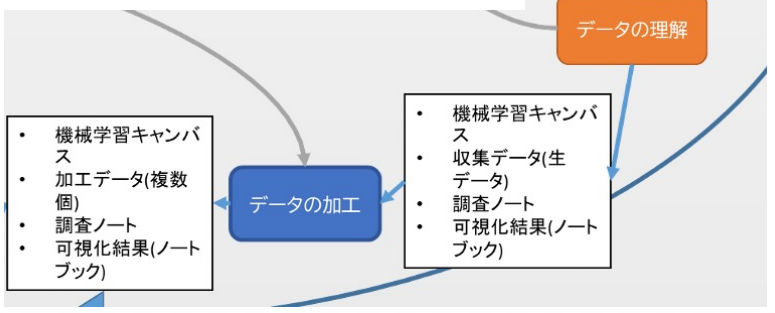
一般的な前処理方法は公開されているが体系立てられていないという問題がある.
体系立てた前処理を行うためには，後続工程とのつながりやPJ全体の進め方を考慮する必要がある.

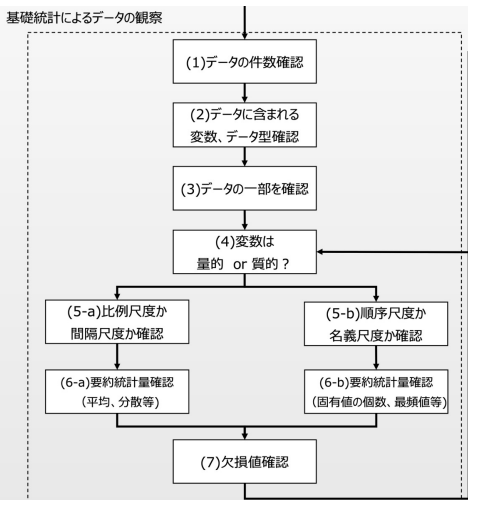
手法・ツールの適用による解決

データサイエンスプロジェクトにおける代表的なプロセスモデルであるCRISP-DMについて，実務での利用を想定した具体化を行い，具体化した工程のうち，データの理解及びデータの加工工程について，作業を体系的かつ円滑に進めるためのフレームワークの検討を行った.

モデル化・アプローチ

・考案したフレームワークの構成要素
  ・改良版CRISP-DM
  ・作業パターン
  ・調査ノート





No	変数名	クレンジング処理
0	pclass	ワンホットエンコーディング
1	name	なし
2	sex	ワンホットエンコーディング
3	age	int64へ変換。小数点以下は切り捨て
4	sibsp	int64へ変換。小数点以下は切り捨て
5	parch	int64へ変換。小数点以下は切り捨て
6	ticket	ワンホットエンコーディング
7	fare	int64へ変換。小数点以下は切り捨て
8	cabin	ワンホットエンコーディング
9	embarked	ワンホットエンコーディング
10	boat	ワンホットエンコーディング
11	body	ワンホットエンコーディング
12	home.dest	ワンホットエンコーディング
13	survived	int64へ変換

評価

フレームワークを利用したAI性能に関する検証

→前処理がバッティングして正当な評価ができず

No	検証内容	スコア(RMSE)
1	すべての工程をAutoMLで実施する場合	0.15031
2	データ加工にはフレームワークを活用し，それ以外をAutoMLで実施する場合	0.15513

初学者を対象としたフレームワークの有効性検証

No	検証内容	Aさんのスコア(accuracy)	Bさんのスコア(accuracy)	Cさんのスコア(accuracy)
1	パターンを利用しない場合	79.6%	80.6%	82.1%
2	パターンを利用した場合	84.4%	82.1%	85.5%

今後の課題

初学者の方に協力いただき効果を確認したが，評価手順を十分に検討することができなかった．正確な評価を実施するためには，フレームワークの活用有無でグループ分けを行い，双方の作業時間やスコアの差を評価する必要があると考える.

＜今後の課題＞
■フレームワークの効果の正確な評価
(1)実務に近い(ダーティな)データを使った検証
(2)評価手順の確立
■フレームワークの改善
(1)作業パターンの改善
(2)調査ノートの改善
(3)2サイクル目以降の詳細化