

# データクレンジング実施時における PUCT を用いた 対象作業選択の最適化に関する一考察

日大生産工(院) ○藤田 直之      日大生産工 水上 祐治

## 1. 研究目的

Ridzuan らによると、データ分析前に実施するデータクレンジング（以下、クレンジング）は、データ分析にかかわるコストの50%近くに達するとされている<sup>1)</sup>。本研究では、クレンジング作業の効率化によるコスト削減を狙い、クレンジングの難易度と担当者の習熟度の関連に着目する。クレンジング作業における達成状況を把握し、効率的な実施計画に基づくクレンジング作業の随時実施を目指す。具体的には、Polynomial Upper Confidence Tree<sup>2)</sup>（以下 PUCT）を適用し、クレンジングの対象作業の選択を漸進的に最適化する方法を提示し検証する。

## 2. 関連研究レビュー

Xuらは、データ分析者は、クレンジング作業において、何らかの決定を下す前に汚れたデータの影響を考慮する必要があることを説明している<sup>3)</sup>。しかし、クレンジングの研究において、戸田らは、多くのクレンジングの研究は、エラー除去の手段として用いられるにとどまっているとしている<sup>4)</sup>。また、戸田らの研究では、プログラム開発を前提として、コードレビューを対象にしているが、プログラム開発が完了したデータ分析前のクレンジング実施に関しては議論をすすめていないとしている<sup>4)</sup>。

## 3. 研究課題

本研究は、データを扱う作業を有する組織において、効率良いクレンジング作業を追究することは業務改善等のコスト低減の観点等で意義があるという立場に立っている。

組織でのクレンジング作業では、日々の作業において、データ作成者とクレンジング担当者の作業が交互に発生する。日々の作業には特徴を有すると考え、以下の仮説を示す。

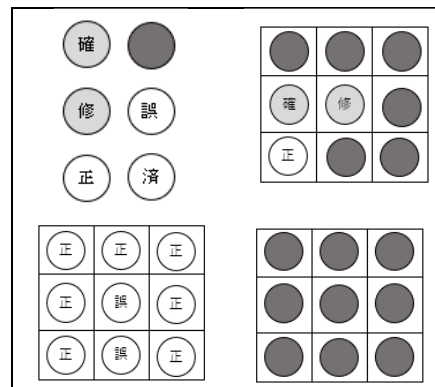


Fig.1 状態を示す駒と盤面

H1: クレンジング結果の精度は、クレンジング作業自体の難易度、データ作成者とクレンジング担当者の組織的習熟度により影響を受ける可能性

H2: データ作成者が有する癖等を含むクレンジングを要するデータの配置には特徴があり、将棋の定石にあたるパターンがある可能性

本研究では、これら仮説の検証のため、クレンジング作業の対象作業の選択を最適化するために定石を浮き彫りにする手法をボードゲームとして提示、検討する。

## 4. 研究方法

クレンジング作業（以下、自軍）とデータ入力者（以下、敵軍）による着手放棄を含む交互に実施する作業を対戦形式でのボードゲームの勝負に置き換える。データセットのマトリックスをボードの升目に見立て、各々のデータを駒と見立てる。駒の状態をFig.1に示す。クレンジング状況の変化は、駒の状態に写像する。今回は、将棋、囲碁と違い、学習の元となる棋譜が無い場合、モンテカルロ木探索を用い、棋譜にあたる学習の元となるパターン（以下、棋譜）を生成する。本研究では、課題解決のアプローチとして、将棋、囲碁の人工知能（以下、AI）で利用されているPUCTを適用し、対象作業選択の最適化を検討

表1 クレンジングゲームのルール

| #  | 内容                              |
|----|---------------------------------|
| 1  | 敵軍は駒「誤」と「正」を配置する                |
| 2  | 敵軍は配置後、駒を全て裏に伏せる                |
| 3  | 自軍と敵軍は交互に駒を捲る（打つ）               |
| 4  | 打つ手が無い時パス可能                     |
| 5  | 裏を捲った時「正」なら「確」に成る               |
| 6  | 裏を捲った時「誤」なら「修」に成る               |
| 7  | 「確」を捲ると「正」に成る                   |
| 8  | 「修」を捲ると「済」に成る                   |
| 9  | 「修」がある時、次はその近傍を打つ               |
| 10 | 「確」「修」がある時、裏は捲れない               |
| 11 | 「正」と「済」の合計数が総駒数の半分以上の時、ゲームを終了する |
| 12 | 終了時、裏を全て捲る                      |
| 13 | 終了時、「誤」があると敵軍の勝ち                |
| 14 | 終了時、「誤」が無いと自軍の勝ち                |

する。クレンジングゲームとしてのルールを表1に規定する。

## 5. 研究実施

PUCTを用いた実験の実施手順を示す。まず、盤面を示す各ノードの対象作業選択の期待値を算出する。対象作業選択の期待値項に未確認項目が選択されやすい状況項を加算し、評価する。(1)に算出する式を示す。次ノードで最大値を選択していく。期待値は、分子に勝ち負けの数、分母にノードの訪問数を示す。勝敗が決定した際に親ノードの期待値を更新する。今回はデータ範囲を3x3に定義し、自軍の算出パターン一例をFig.2に示す。評価値を基に対象選択の候補を確認した。

## 6. 考察

クレンジング作業をボードゲームに見立て、PUCTを適用し、評価値を算出する仕組みを検討し、クレンジング対象の選択が可能か確かめた。早めにクレンジングが完了すると期待値が上るので、実際の作業においてクレンジングの成功確率が向上するのではと考えている。今後は、選択精度を向上するために、棋譜を増やすこと、データ範囲を拡張すること、誤データから修正済データを1元で決定

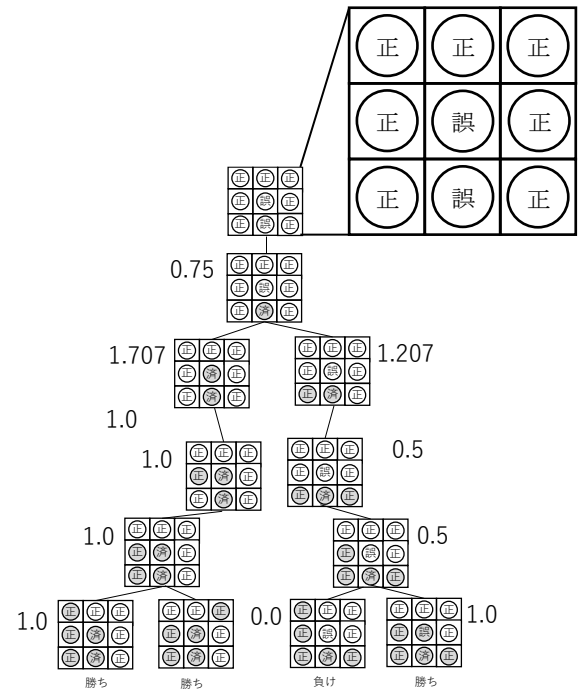


Fig.2 算出パターンの例

しないで、様々な修正情報、例えば、慣習による日本語のブレ等による誤入力等に対応するためにルールを拡張することを検討する。

## 参考文献

- 1) Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon, "A Review on Data Cleansing Methods for Big Data", Procedia Computer Science, vol.161, pp.731-738 (2019)
- 2) David Auger, Adrien Couetoux, Olivier Teytaud. Continuous Upper Confidence Trees with Polynomial Exploration - Consistency. ECML/PKDD 2013, Sep 2013, Prague, Czech Republic. Springer Verlag, 8188, pp.194-209, LNCS. (2013)
- 3) Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, Jiannan Wang: Data Cleaning: Overview and yEmerging Challenges, SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data, pp2201-2206 (2016)
- 4) 戸田航史, 亀井靖高, 吉田則裕: コードレビュー分析におけるデータクレンジングの影響調査, 情報処理学会論文, Vol. 58, No. 4, pp. 845-854 (2017)

$$f(x) = \bar{x}_i + \sqrt{\frac{2 \log n}{n_i}} \quad (1)$$