LightGBMによる
書き手の感情極性分類タスク

人工知能研究室 B3 梶川怜恩

2023 年 1 月 11 日

タスク

- 日本語のTwitterテキストについての感情極性分類
- 書き手の感情極性を5クラス分類（-2, -1, 0, 1, 2）
- 評価指標：Quadratic Weighted Kappa

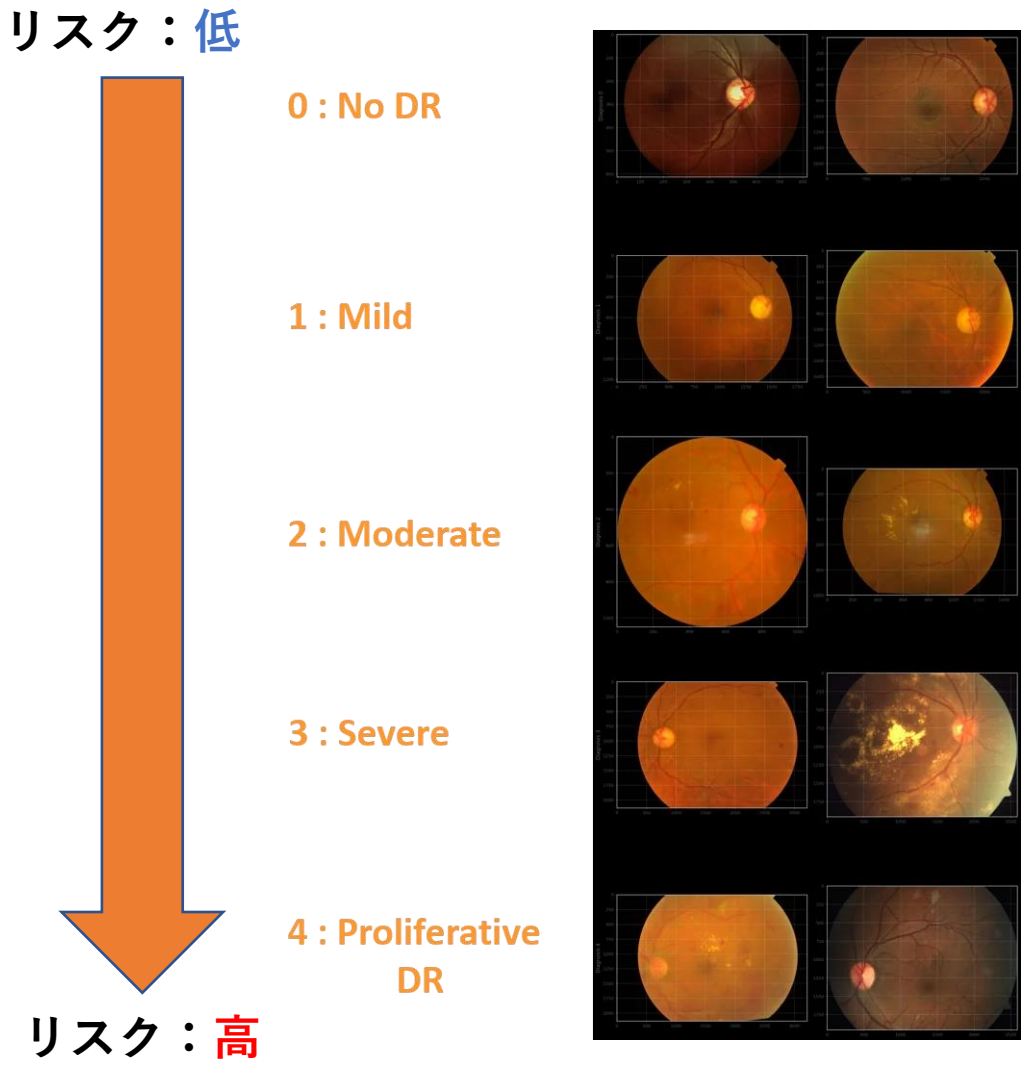
- 分割の変更なし
- ニューラルネットワーク使用禁止・外部データを使用しない

学習	評価	提出
30,000	2,500	2,500

Quadratic Weighted Kappa(QWK)

- マルチクラス分類用の評価指標
- クラス間に順序関係
- ラベル分布に大きく影響する
- 予測を大きく外すほどペナルティ

数字が大きくなるにつれてリスクが増す
＝クラス間に順序関係

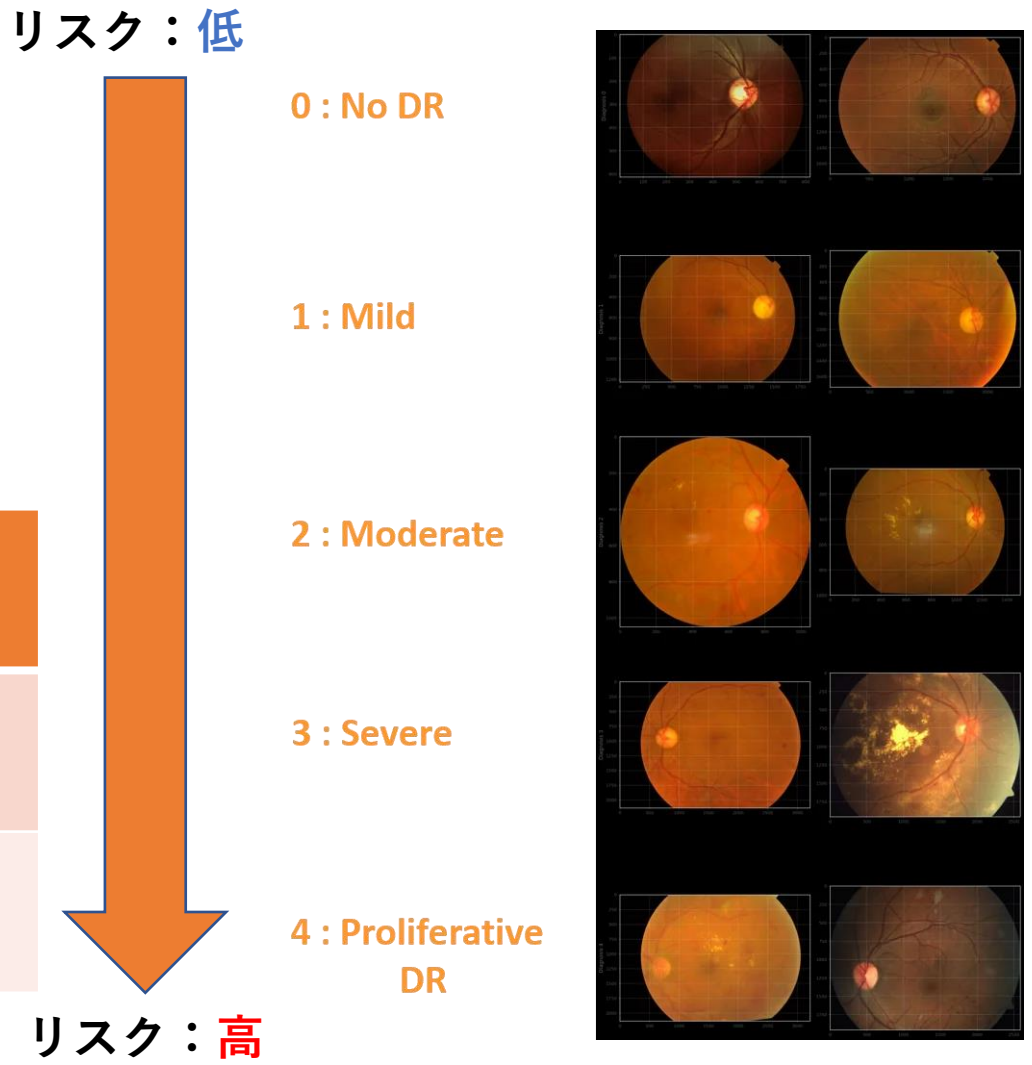


[1] https://www.kaggle.com/competitions/aptos2019-blindness-detection

Quadratic Weighted Kappa(QWK)

- マルチクラス分類用の評価指標
- クラス間に順序関係
- ラベル分布に大きく影響する
- **予測を大きく外すほどペナルティ**

正解	0	2	4	ACC	QWK
予測1	0	2	1	0.66	0.0
予測2	0	1	3	0.33	0.85



アイデア概要

①前処理・特徴量生成

②LightGBMによる回帰タスクへの帰着

③アンサンブル

工夫１：前処理・特徴量生成

- 単語分割
  - Neologdn による文正規化
  - Sudachi-A – small dict 、sudachiによる単語正規化
  - 数字の正規化(任意の数字→0)

- TF-IDFによる文ベクトルの生成
  - TfidfVectorizerによる生成
  - 特徴量数を3100
  - 文書ベクトルのノルムはL1正則
  - 文字単位で特徴量を生成

工夫２： LightGBMによる回帰タスクへの帰着

- 順序関係があるクラス分類タスク
- 回帰問題として解くことが定石
- 閾値の設定
  - OptimizedRounder[2]による閾値の最適化→汎化性能○
    →今回は、提出用のデータ分布にだけ合った閾値
    →ヒューリスティックな手法にならざるを得ない

- 過去のコンペ事例を活用したハイパラの採用[3]

[2] https://www.kaggle.com/c/petfinder-adoption-prediction/discussion/76107#latest-502207

[3] https://github.com/nyanp/nyaggle/blob/master/nyaggle/hyper_parameters/lightgbm.py

工夫３：アンサンブル

- 多数決による
- 異なるハイパーパラメータでの予測値を採用

結果・比較

- LightGBMに各手法を追加したスコアの比較

モデル	QWK*100
LightGBM + OptimizedRounder	44.9
LightGBM + 閾値の調整	50.4
LightGBM + 閾値の調整 + ensemble	**51.0**

没ネタ(コスト・ルールNG)

- アンサンブルへの工夫
  - 多数決
  - 重み付け
- 別モデルとのアンサンブル
  - ロジスティック回帰
  - サポートベクター回帰
- データセットの校正
  - ライブラリでは除去できない誤字脱字の人手校正
- 顔文字の除去
  - nagisa(形態素解析器)の利用[4]
    →NNで学習していたため不採用

[4] https://qiita.com/dcm_murakami/items/4c016936a739bfb2a517

感想

- 一つのモデルに対して改良を行えた
- 閾値の調整にかなり時間がかかった
- もう少し余裕をもって取り組みたい