擬似ラベルを用いた半教師あり学習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 2 月 15 日

タスク

- 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
- 外部データを使用しない
- **PyTorchによるNNの設計**
- **Testデータに対するハイパラ調整なし**

Train	Valid	Test
30,000	2,500	2,500

アイデア概要

①前処理・特徴量生成

②擬似ラベルによる半教師あり学習

③モデル構造

工夫１：前処理・特徴量生成

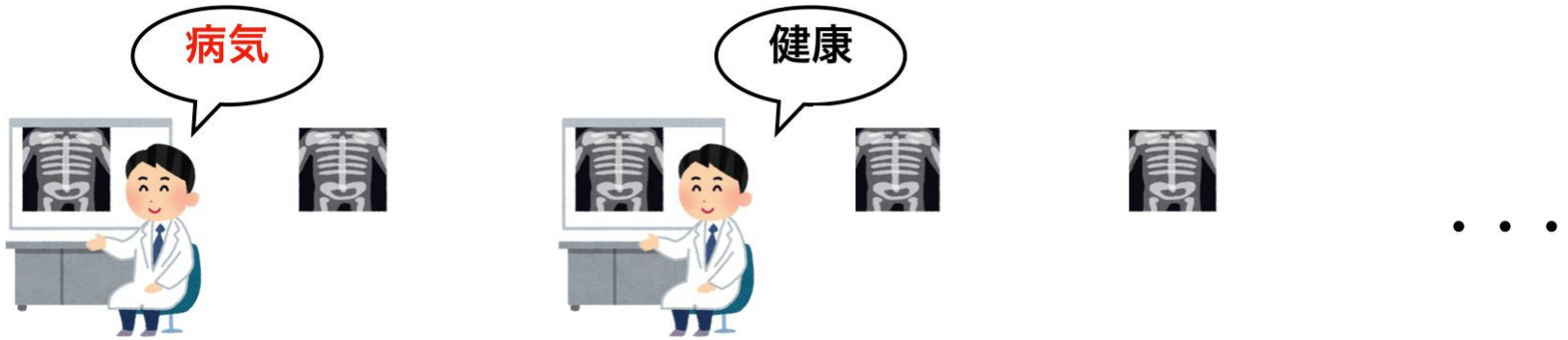
- 単語分割
  - Sudachiによる単語正規化
  - 分割モード・辞書の設定
  - 数字の正規化(任意の数字→0)

- TF-IDFによる文ベクトルの生成

工夫２：疑似ラベルによる半教師あり学習

半教師あり学習

- ラベル付けされたデータと、**未ラベルデータ**を使って学習
- ラベル付きデータが少ない状況に適している



擬似ラベル(Pseudo-Label)

- 未ラベルデータに対するラベル付けの方法の１つ

1. ラベル付けされたデータで学習済みモデルを作成
2. 1.のモデルで未ラベルデータを推論
3. 推論したラベルを**擬似ラベル**とし、ラベル付けされたデータに混ぜて学習させる

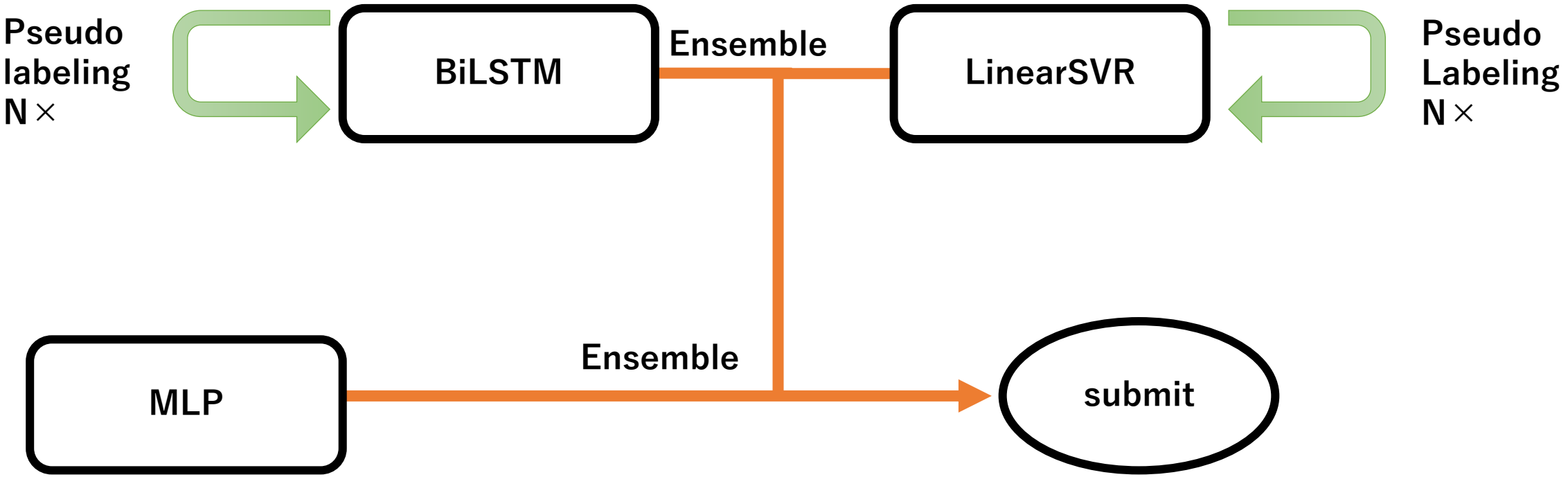
→**Testデータ**に対して適応

工夫３： モデル構造

- Bidirectional LSTM(双方向LSTM)
  - 擬似ラベル適応
  - 隠れ層の中間の位置と最終位置の平均ベクトル
- LinearSVR
  - 擬似ラベル適応
  - Optuna による閾値最適化
- MLP
  - 一般的な分類タスク

工夫３： モデル構造

Testデータに対して擬似ラベルを適用



結果・比較

- 擬似ラベルによる性能向上を確認

モデル	QWK
BiLSTM(simple)	0.455
BiLSTM(Pseudo-Label)	**0.469**

- QWK: 多クラス分類における評価指標(-1.0~+1.0)

結果・比較

- スコアの比較

モデル	QWK
BiLSTM(simple)	0.455
BiLSTM(Pseudo-Label)	0.469
BiLSTM + LinearSVR(optuna)	0.522
BiLSTM + LinearSVR(optuna) + MLP	**0.534**

没ネタ

- TextCNN
  - QWKが低い
- QWK-lossの実装
  - lossがNaNになる
  - 自作の損失関数の設計ができるようにしたい

まとめ

- 採用した手法による改善を確認できた
- 体調管理に気を付ける