Poster Presentation

May 2023

VICTORINO John Noel

2023年5月23日

Overview for May

- **ABC 2023 Paper:** Handling Class Imbalance with Non-Uniform Negative Sampling, Log Odds Correction, and Oversampling in Forecasting Parkinson's Disease Wearing-off with Fitness Tracker Dataset
    - Paper: In Progress
    - Code & Results: In Progress
    - Presentation: for May 2023 Poster Presentation

Proposed Method

Forecasting Wearing-off using **Negative Sampling and Log Odds Correction**
1. Train "pilot model" using uniform sampling of both classes with equal sample sizes.
1. Score data using pilot model.
2. Sample data using **Negative Sampling**.
3. Correct likelihod using **Log Odds Correction**.

- **Contributions**
  - Application of Negative Sampling and Log Odds Correction to assign larger probabilities to more informative instances, i.e., wearing-off labels and informative normal labels.
  - Improves forecasting of wearing-off by *X%* as compared to ABC 2022 paper.
- **Paper category**: Application-based paper

Forecasting Wearing-off Task

- **Wearing-off** is a phenomenon where the effect of a drug wears off before the next dose is taken.
- Let $\{(X_t, y_t)\}_{t=1}^{N}$ be the training data that satisfies

$$Pr(y_{t+1} = 1|M) = p(M;\theta), M = \begin{bmatrix} X_t, & y_t \\ \vdots & \vdots \\ X_{t-1\text{hour}}, & y_{t-1\text{hour}} \end{bmatrix}$$

- $y \in \{0, 1\}$ is wearing-off (1) or not (0),
- M is the feature matrix of X and y from the last hour until time t,
- θ is the model parameters

Forecasting Wearing-off Task

For example, if the dataset is in 15-minute interval.

- Forecast: y_{t+4}
- Input:

$$M = \begin{bmatrix} X_t & y_t \\ X_{t-1} & y_{t-1} \\ X_{t-2} & y_{t-2} \\ X_{t-3} & y_{t-3} \\ X_{t-4} & y_{t-4} \end{bmatrix}$$

Imbalance in Wearing-off Dataset

Wearing-off Statistics

- Mean: 8.276%
- Median: 7.682%
- Standard Deviation: $\pm$ 3.624
- Participants Below Median: 1, 2, 7, 9, 12, 13
- Participants Equal / Above Median: 3, 4, 5, 6, 8, 10, 11

Participant	Normal ($y = 0$)	Wearing-off ($y = 1$)
1	92.361	7.639
2	97.348	2.652
3	86.771	13.229
4	90.492	9.508
5	92.318	7.682
6	89.815	10.185
7	98.821	1.179
8	87.973	12.027
9	92.992	7.008
10	86.771	13.229
11	90.705	9.295
12	92.519	7.481
13	93.528	6.472

Imbalanced Data Definition

- Let N_1 wearing-off cases, N_0 normal cases, and $N = N_1 + N_0$ total cases.
- When N_1 much smaller than N_0, appropriate to assume that N_1 increases in slower rate compared with N_0, i.e.,

$$\frac{N_1}{N_0} \xrightarrow{P} 0 \text{ and } N_1 \xrightarrow{P} \infty \text{ as } N \to \infty$$

- This requires that $Pr(y_{t+1} = 1) \to 0$ as $N \to \infty$ on model side.

Imbalanced Data Definition

Assume that $\theta = (\alpha, \beta^T)^T$, log odds is as follows:

$$g(M;\theta) := \log\{\frac{p(M;\theta)}{1-p(M;\theta)}\} = \alpha + f(M;\beta)$$

- $f(M;\beta)$ is the smooth function of β, e.g., neural network or logistic regression.
- The true parameter $\theta^* = (\alpha*, {\beta^*}^T)^T$ with assumption that
  - $\alpha^* \to -\infty$ as $N \to \infty$,
  - β^* is essentially fixed, as long as β^* has a finite limit
- For both assumptions indicate that marginal & conditional probabilities of positive instance are **small**.
- This means that changes in covariates do not convert small probability event to large probability event.

Number of Positive Instance Matter

- For rare event, estimation error rate for θ is related to N_1 instead of N.
- Using the full training data, maximum likelihood estimator (MLE) of θ is

$$\hat{\theta}_{MLE} := \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \{y_{t+1_i} \cdot g(M_i;\theta) - \log(1 + e^{g(M_i;\theta)})\}$$

(1) Negative Sampling

- **Keep all wearing-off instances while significantly subsampling negative instances**, since available information ties with wearing-off instances instead of full training data.
  - **Uniform sampling**: keeping the ratio of positive to negative samples.
  - Reduce the number of negative samples using **nonuniform sampling**.

(1) Negative Sampling

Algorithm 1: Negative Sampling

For $i = 1, \ldots, N$:

  if $y_{t+1_i} = 1$, include $\{M_i, y_{t+1_i}, \pi(M, y_{t+1}) = 1\}$ in the sample.

  if $y_{t+1_i} = 0$, include $\{M_i, y_{t+1_i}, \pi(M, y_{t+1}) = \rho\varphi(M)\ \}$ in the sample

    if $u_i \leq \rho\varphi(M)$, where generate $u_i \sim \mathbb{U}(0, 1)$.

- ρ: sampling rate on negative instances.
- $\pi(M, y_{t+1})$: sampling probability for given (M, y_{t+1}).
  - $\pi(M)$: sampling probability for negative cases, as shorthand.
- $\varphi(M)$: "pilot model" from uniform sampling of both classes with equal sample sizes.
- $\delta = 1$ if data point is selected; otherwise, $\delta = 0$.

(2) Log Odds Correction

Recap...

$$Pr(y_{t+1} = 1|M) = p(M;\theta) = \frac{1}{1 + e^{-g(M;\theta)}},$$

$$\hat{\theta}_{MLE} := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \{y_{t+1_i} \cdot g(M_i;\theta) - \log(1 + e^{g(M_i;\theta)})\}$$

For the **Log Odds Correction**, for the data included in the subsample (where $\delta = 1$), the conditional probability of $y_{t+1} = 1$

$$Pr(y_{t+1} = 1|M, \delta = 1) = \frac{1}{1 + e^{-g(M;\theta) - l}}, \text{where } l = -\log\{\pi(M)\}$$

$$\hat{\theta}_{lik} := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \delta_i [y_{t+1_i} \cdot g(M_i;\theta) - \log\{1 + e^{g(M_i;\theta) + l_i}\}]$$

This is done to

- avoid assigning smaller weights to more informative instances
- avoid biased estimator due to Algorithm 1

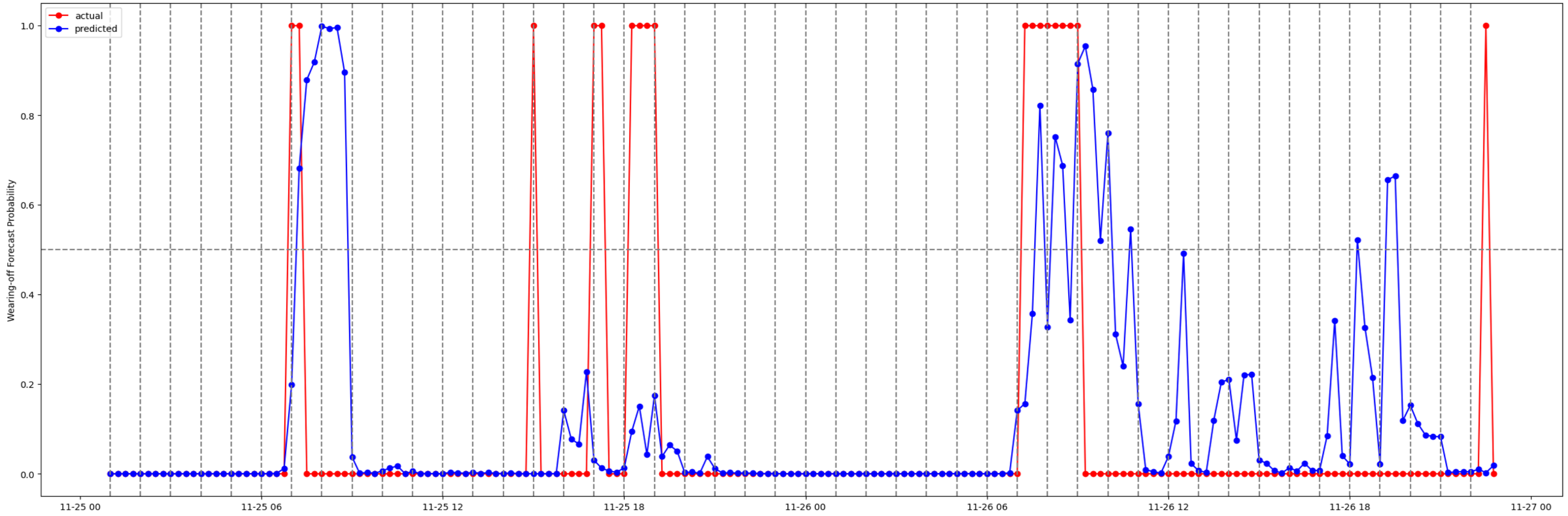
Summary of Method

Forecasting Wearing-off using **Negative Sampling and Log Odds Correction**

1. Train "pilot model" using uniform sampling of both classes with equal sample sizes.
1. Score data using pilot model.
2. Sample data using **Algorithm 1: Negative Sampling**.
3. Correct likelihod using **Log Odds Correction**.

Initial Results

- Still in Progress: Because of wrong coding.
- Baseline Model using XGBoost




```
f1 score: 0.27027027027027023
recall: 0.2777777777777778
accuracy: 0.8532608695652174
              precision    recall  f1-score   support

         0.0       0.92      0.92      0.92       166
         1.0       0.26      0.28      0.27        18

    accuracy                           0.85       184
   macro avg       0.59      0.60      0.59       184
weighted avg       0.86      0.85      0.86       184
```

- The idea of the proposed method is to:
    - Reduce negative samples and as a result improve the performance of the model.
    - Boost probabilities from positive samples.

Next Steps

- Finish ABC paper (if forecast results have improved.)
- Otherwise, try to improve forecast results using other methods previously studied.
  - Then, submit to journal.
- Prepare data collection using new Garmin App Collector to include
  - Accelerometer
  - HRV
- Development work on Pose Estimation
- Work on Additional tasks due to ABC 2023 challenge