Striking a Balance: Co-Training Framework for Enhancing Survey Accuracy While Reducing Respondent Burden in Travel Data Collection

AlOlabi Reem, Makoto Chikaraishi*
Graduate School of Advanced Science and Engineering, Hiroshima University, Japan
d232824@hiroshima-u.ac.jp, chikaraishim@hiroshima-u.ac.jp
* Corresponding author

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

March 8, 2025

Keywords: Co-training, Semi-supervised learning, Travel mode detection, Respondent burden, Data accuracy

1 INTRODUCTION

Conventional travel diaries are invaluable for analyzing travel behaviors but are limited by high respondent burden, recall bias, and extensive manual labeling, which reduce data quality and scalability. Advances like GPS tracking and mobile apps have streamlined data collection but still lack the contextual labeling needed for detailed travel analysis, especially for identifying transportation modes. This study addresses these gaps by introducing a novel co-training framework that integrates data collection and model development in a unified process, redefining traditional survey boundaries. Unlike conventional methods, which separate data collection and labeling, our framework embeds co-training within the data collection phase, using labeled and unlabeled GPS data to iteratively enhance model accuracy.

Our research aims to improve survey efficiency and data accuracy by applying semi-supervised learning to real-time transportation mode detection. Specifically, we evaluate the impact of varying labeled-unlabeled data ratios, optimizing accuracy with minimal manual labeling. Tested on travel data from Hiroshima, Japan, our framework's performance is compared with standard supervised models, demonstrating significant potential for scalable, low-burden travel surveys. This unified approach advances data collection methodologies by dynamically balancing respondent burden with data quality, offering a resource-efficient solution for transportation research and beyond.

2 Methodology

This study proposes a co-training framework designed to improve transportation mode detection in settings where labeled data is scarce but unlabeled data is plentiful—a common scenario in large-scale travel surveys. By integrating labeled and unlabeled data within a unified, iterative framework, this approach minimizes dependence on manual labeling while significantly enhancing the accuracy and efficiency of traditional survey methods.

2.1 Co-Training Algorithm: Theoretical Foundations and Design

The co-training algorithm, introduced by Blum and Mitchell (1998), is a semi-supervised learning method suitable for scenarios with abundant unlabeled data, such as GPS-based transportation surveys. In this framework, two classifiers are trained on distinct views of the data, iteratively exchanging high-confidence pseudo-labels to refine their predictive accuracy. In this study, **view 1** contains speed-related features and **view 2** incorporates GIS-based, weather, and socio-economic features, such as proximity to transit stops and local demographic characteristics. The classifiers exchange pseudo-labels, allowing each classifier to improve through feedback from the complementary view, effectively increasing the accuracy of each classifier without requiring additional labeled data.

2.2 Dataset Structure and Terminology

The dataset used for co-training framework, referred to as **DTrip data**, is collected via a mobile travel diary application, combining labeled data actively collected from participants with passively collected GPS data. This dataset is divided into:

- **Labeled Data (L)**: Consisting of trips with verified transportation modes.
- **Unlabeled Data (U)**: Consisting of trips with unknown transportation modes, which are progressively pseudo-labeled through the co-training process.

The unlabeled data are subdivided into **pseudo-labeled data (U^P)** and **non-labeled data (U^N)** as the iterative labeling progresses. The overall structure of the data used in co-training is illustrated in **Figure 1**.

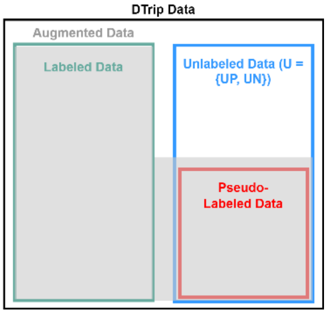

Figure 1: Composition of Data Used in the Co-Training Framework

2.3 Co-Training Algorithm Design

The co-training process begins with initializing two classifiers, each trained on its respective labeled dataset subset. During each iteration:

1. Each classifier generates predictions on the unlabeled data and identifies instances with high-confidence predictions, defined by a preset confidence threshold (e.g., 85%).
2. These high-confidence predictions are shared as pseudo-labels with the other classifier, expanding the labeled dataset.
3. The process iterates, refining each classifier's predictions through feedback from the other classifier until either no new high-confidence labels are produced, or a maximum iteration count is reached.

The algorithm design includes a filtering step (see **Figure 2**), where only high-confidence predictions are incorporated to ensure data quality. This iterative pseudo-label exchange and selective filtering enhance model robustness by preventing error propagation.

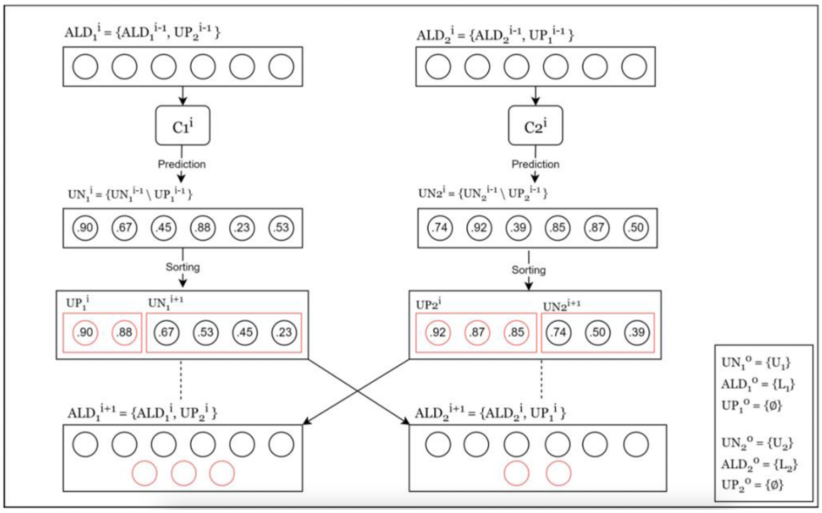

Figure 2: Co-Training Process for Semi-Supervised Learning

2.4 Integrating Co-Training in Survey Scheme

This methodology integrates the co-training framework directly into the survey process, creating an adaptive system that continuously refines model performance by iteratively combining labeled and pseudo-labeled data. Unlike conventional methods that treat labeled and unlabeled data separately, this framework allows for real-time model refinement by alternating between **Model Refinement** and **Filtering Unlabeled Data**, as illustrated in Figure 3. In each iteration, two models trained on distinct feature views exchange high-confidence pseudo-labels to enhance prediction accuracy. The high-confidence pseudo-labeled data is then added to the labeled dataset (L), expanding the model's knowledge base without additional manual labeling. The iterative process produces two refined models and an **Augmented Labeled Dataset (ALD)**, which combines both the original labeled data (L) and the pseudo-labeled data (U^P). This final output can be applied in two ways:

- **Method 1**: Select the best-performing classifier at the end of co-training and use it directly for labeling new instances.
- **Method 2**: Use the augmented dataset (ALD) to train a new, comprehensive mode detection model.

This dual-purpose framework supports both **Data Enrichment**—by building a robust labeled dataset through pseudo-labeling—and **Model Refinement**—by selecting the most accurate classifier based on performance metrics such as F1 score. This approach reduces respondent burden, minimizes manual labeling, and adapts dynamically to incoming data, providing a scalable and resource-efficient solution for large-scale transportation mode detection.

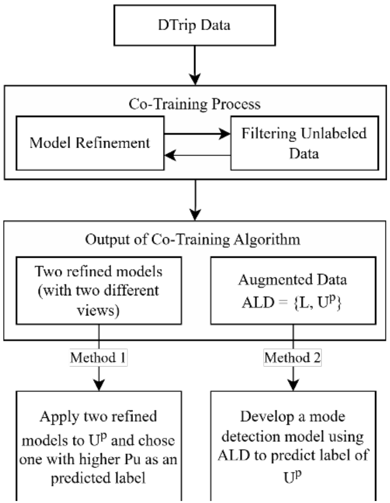
DTrip Data
Co-Training Process
Model Refinement
Filtering Unlabeled Data
Output of Co-Training Algorithm
Two refined models (with two different views)
Augmented Data ALD = {L, U^P}
Method 1
Method 2
Apply two refined models to U^P and chose one with higher Pu as an predicted label
Develop a mode detection model using ALD to predict label of U^P

Figure 3: Co-Training Workflow and Data Augmentation in Transportation Mode Detection

3 Data and Experimental Design

Data was collected from three smartphone-based travel diary surveys in Hiroshima, Japan, covering three seasons (December 2018–January 2019, January–February 2020, and October–November 2020), resulting in over 12,000 trips with GPS trajectories, socio-economic details, and weather conditions to support accurate mode detection. The data underwent cleaning and preprocessing as shown in **Figure 4** (removing missing timestamps and outliers like speeds over 150 km/h), trip segmentation (using a 3-minute threshold to define boundaries), and noise reduction (applying a moving average filter to smooth erratic GPS data, especially for walking). Key features, such as speed metrics (e.g., average speed, acceleration) and contextual information (e.g., proximity to transit stops, weather, and demographic factors like age and occupation), were selected to capture travel dynamics and context, ensuring comprehensive and reliable data for transportation mode classification.

To evaluate the effectiveness of the co-training framework, we implemented a series of experiments across nine distinct scenarios that varied the proportions of labeled and unlabeled data. These scenarios, ranging from 90% labeled data in Scenario 1 to 90% unlabeled data in Scenario 9, were designed to simulate real-world constraints in data availability and assess the optimal balance for model performance (see Figure 7). Additionally, we employed multiple feature configurations, including cases that isolated speed metrics, combined GIS and weather data, and grouped contextual information to understand how these feature views influenced co-training outcomes. Resampling techniques were applied to ensure balanced representation across transportation modes, allowing for more accurate classification and reducing the model's bias toward dominant modes like driving and walking. Comparative analyses were conducted between the co-training framework and traditional supervised models (SVM, KNN, NN, and RF). These models were evaluated with and without the augmented datasets generated by co-training, highlighting the impact of pseudo-labeling on performance under data-scarce conditions.

5 Results

The co-training framework demonstrated significant improvements in classification accuracy across all tested scenarios, particularly in data-limited contexts where labeled data was scarce. Key findings from the experimental analysis include:

- **Optimal Data Balance for Co-Training**: The experiments across nine scenarios revealed that a moderate balance of labeled and unlabeled data (Scenario 3, with 70% labeled data and 30% unlabeled data) yielded the best performance, achieving an accuracy of 87.1%. This balance allows the model to fully leverage pseudo-labeling without over-relying on manual labels, thus reducing labeling costs while maximizing predictive accuracy.

- **Effectiveness of Feature Configurations**: The framework's success also varied depending on the feature configuration used. The configuration that separated speed-based features (S) from contextual data (GIS-based transit proximity, weather, and socio-economic features) yielded the highest accuracy. This arrangement allowed each classifier to leverage complementary data perspectives, particularly improving detection of travel modes with distinct contextual signatures, such as public transit and cycling.

- **Comparative Advantage over Baseline Models**: The co-training framework showed a notable performance advantage over traditional supervised models (SVM, KNN, NN, and RF), especially in scenarios with limited labeled data. When pseudo-labeled data generated by the co-training process was used to augment the training dataset and then train the conventional algorithm using it, the algorithms demonstrated a significant improvement comparing to when they were trained only on the labeled data, This comparative analysis confirms that the co-training framework, with its iterative pseudo-labeling, effectively leverages unlabeled data to enhance classification accuracy, offering a practical and efficient alternative to models relying exclusively on large, manually labeled datasets.

4 Conclusion

This study introduces an integrated co-training framework to address the challenges of manual labeling in travel surveys, particularly the limitations posed by high respondent burden and recall bias in traditional data collection methods. By leveraging both labeled and unlabeled GPS data in an iterative, semi-supervised learning process, the co-training framework demonstrates substantial improvements in classification accuracy and resource efficiency. The experimental findings highlight that a balanced mix of labeled and unlabeled data optimizes performance, offering a practical solution for large-scale data-limited environments. Additionally, comparisons with baseline models confirm the framework's robustness, effectively utilizing pseudo-labeled data to achieve high accuracy in transportation mode detection with minimal manual intervention. This framework not only advances the field of travel behavior research but also provides a scalable, low-burden approach adaptable to broader applications in urban planning and data-driven policy development.

References

Blum, A.and Mitchell, T. Combining labeled and unlabeled data with co-training. Proceedings of Proceedings of the eleventh annual conference on Computational learning theory, pp. 92-100, 1998.