Semi-Supervised Learning

World Learning
EDUCATION | DEVELOPMENT | EXCHANGE

Start

Introduction.exe

Semi-Supervised Learning (SSL) is a machine learning paradigm that bridges the gap between supervised and unsupervised learning by leveraging both labeled and unlabeled data. Since labeled data is scarce and costly, while unlabeled data is abundant, SSL maximizes learning efficiency with minimal supervision. Used in fields like image recognition and NLP, it enables AI to train smarter and faster through techniques like self-training and consistency regularization, making powerful models possible even with limited labeled data.

SelfTrainingMethod.exe

Self-training is a popular Semi-Supervised Learning technique where a model learns iteratively by labeling its own unlabeled data. It starts with a small set of labeled data, makes predictions on unlabeled samples, and adds the most confident predictions to the training set. This process gradually improves the model's accuracy, making it a simple yet effective way to leverage unlabeled data for better learning.

AI Forge Program 2025

Riham Belal

Chahrazed Bousmaha

Mohamed Nazim Hamza

Fatima Zohra Homrani

Naila Lamia Bettouche

Pros.exe

- Time and money saving: Labeling data is time consuming and expensive and often requires domain expertise, so using a small set of labeled data and a large set of unlabeled data can be cheaper and faster.
.
- Improved model performance: In many cases, semi-supervised learning models can achieve better accuracy compared with models trained only on labeled data, especially when labeled data is scarce.

- Effective for unstructured data: Semi-supervised learning is particularly well-suited for tasks such as text, video, or audio categorization, where unlabeled data is often abundant.

- Works Well with Natural Data: SSL algorithms thrive in complex real-world datasets like text, images, and audio, where labeling every sample is impractical.

Yes

KeyConcepts.exe

- Datasets: Small labeled dataset (training data) and a large unlabeled dataset(testing data)which should be diverse .The split depends on your specific needs.

- Initial Training: Train a classification algorithm using the labeled training data.

- Pseudo-labeling: Use the trained classifier to predict labels for the unlabeled data instances. Some of these predicted labels, the ones with the highest probability of being correct are adopted as "pseudo-labels", to select these pseudo-labels, it is crucial to set a confidence threshold which minimizes the risk of adding wrong pseudo-labels. All confident model predictions are converted into “one-hot” vectors, where the most confident class becomes the label. From this, we train on the new “one-hot” probability distribution as a pseudo-label.

- Dataset Augmentation: Concatenate the pseudo-labeled data with the labeled training data. Re-train the classifier on the combined pseudo-labeled and labeled training data.

- Evaluation: Use the trained classifier to predict class labels for the labeled data. Then evaluate the classifier performance using your metric of choice.

- Iterate and refine: This loop continues until performance stabilizes or reaches a desired level.

Cons.exe

- Sensitive to labeled data quality: The accuracy and relevance of labeled data significantly affects the model's performance, so care and money needs to be allocated to ensure quality labeling.

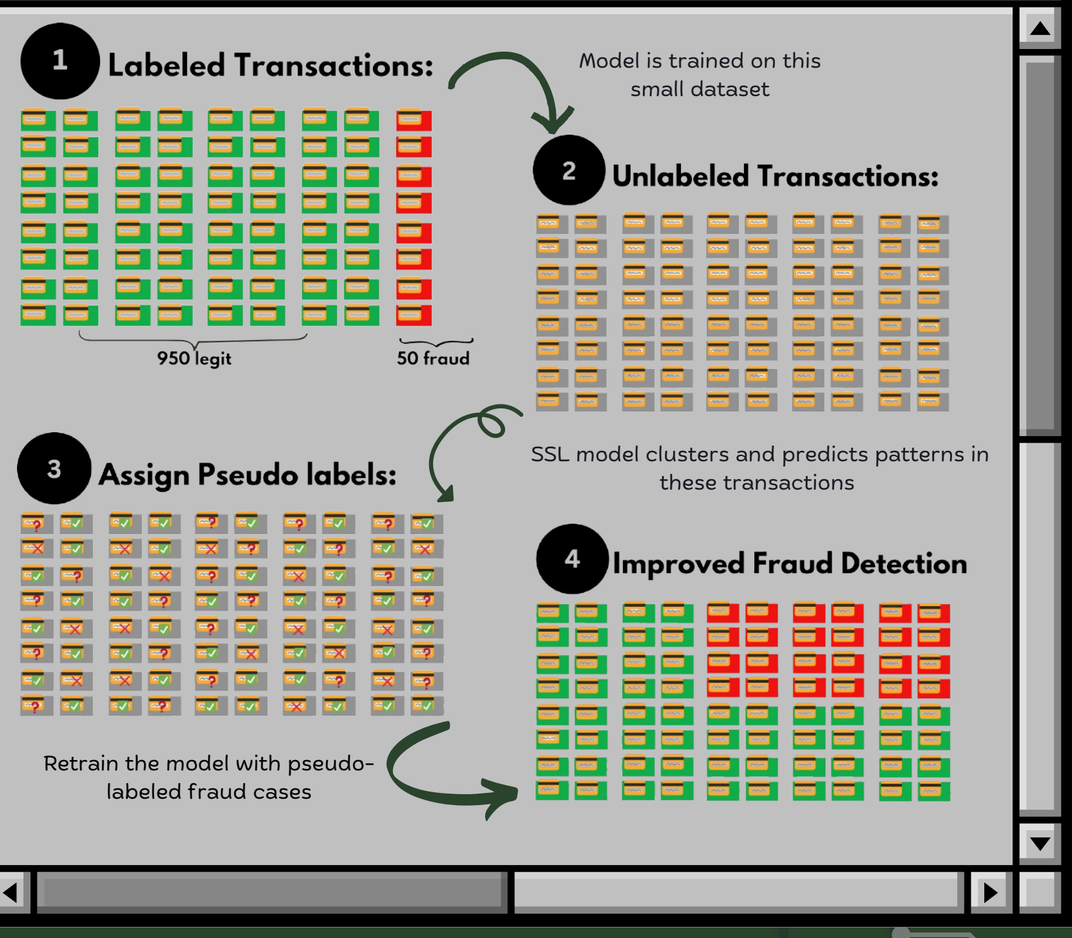
- Error Amplification: Incorrect pseudo-labels can introduce noise and reinforce mistakes, especially if the model confidently labels data incorrectly during early iterations.

- Limited Algorithm Choices: Not all machine learning algorithms are easily adaptable to semi-supervised learning, and some require significant customization.

- Computational complexity: Handling large datasets with iterative retraining can get expensive.

No

Fraud.png



RealWorldApplication.exe

Financial institutions face challenges in detecting fraud due to the limited number of labeled fraudulent transactions (only 0.1% - 0.5% of transactions are fraudulent). Semi-supervised learning helps by combining a small set of labeled fraud cases with a large amount of unlabeled transactions to identify hidden patterns. The model first learns from known fraud cases, then analyzes unlabeled transactions, assigning pseudo-labels to suspicious ones. Over time, the system retrains itself with these new insights, improving accuracy while reducing false positives. This approach enables banks to detect fraudulent activities more efficiently, enhancing security and preventing financial losses.

Cancel