Confidence Thresholding in Self-Training: A Tutorial

Semi-Supervised Learning

Author: Md Salahuddin Chowdhury
23029856

What We Cover:

- Topic We Cover
- What is Self-Training?
- Dataset Visualization
- Key Observations from the Dataset
- Initial Model Training
- Introducing Confidence Thresholding
- Retraining the Model
- Results and Analysis
- Conclusion
- References

Topic We Cover

In machine learning, one of the significant challenges is learning from limited labeled data. Semi-supervised learning provides a solution by leveraging unlabeled data alongside labeled data. Among semi-supervised methods, self-training is widely recognized for its simplicity and effectiveness. However, self-training can suffer from noisy pseudo-labels, which degrade model performance. This tutorial aims to address this issue by demonstrating the use of confidence thresholding, a technique that filters pseudo-labels based on their confidence levels, to improve self-training outcomes.

This tutorial is structured to:

- Provide an overview of self-training and semi-supervised learning.
- Demonstrate confidence thresholding in self-training using visual explanations.
- Analyze the results and offer practical insights into applying this method.

What is Self-Training?

Definition of Semi-Supervised Learning

Semi-supervised learning is a type of machine learning that leverages a small amount of labeled data and a large amount of unlabeled data to improve model performance. It bridges the gap between supervised and unsupervised learning (Zhu, 2008).

What is Self-Training?

Definition of Self-Training

Self-training is an iterative semi-supervised learning method where a model predicts pseudo-labels for unlabeled data and retrains itself using both labeled and high-confidence pseudo-labeled data (Lee, 2013).

What is Self-Training?

Why Self-Training?

Self-training is simple, flexible, and can work with most supervised learning models. However, its success depends on the quality of pseudo-labels, which can be enhanced through techniques like confidence thresholding.

What is Self-Training?

Semi-Supervised Learning

- Combines small labeled datasets and large unlabeled datasets.
- Bridges the gap between supervised and unsupervised learning (Zhu, 2008).

Self-Training

- Iterative process:
    - Train a model on labeled data.
    - Predict pseudo-labels for unlabeled data.
    - Use high-confidence pseudo-labels to retrain.
- Simple and flexible but depends on pseudo-label quality (Lee, 2013).

Dataset Visualization

Setup

- Small percentage of labeled samples.
- Large pool of unlabeled samples.
- Goal: Use unlabeled data to enhance model performance.

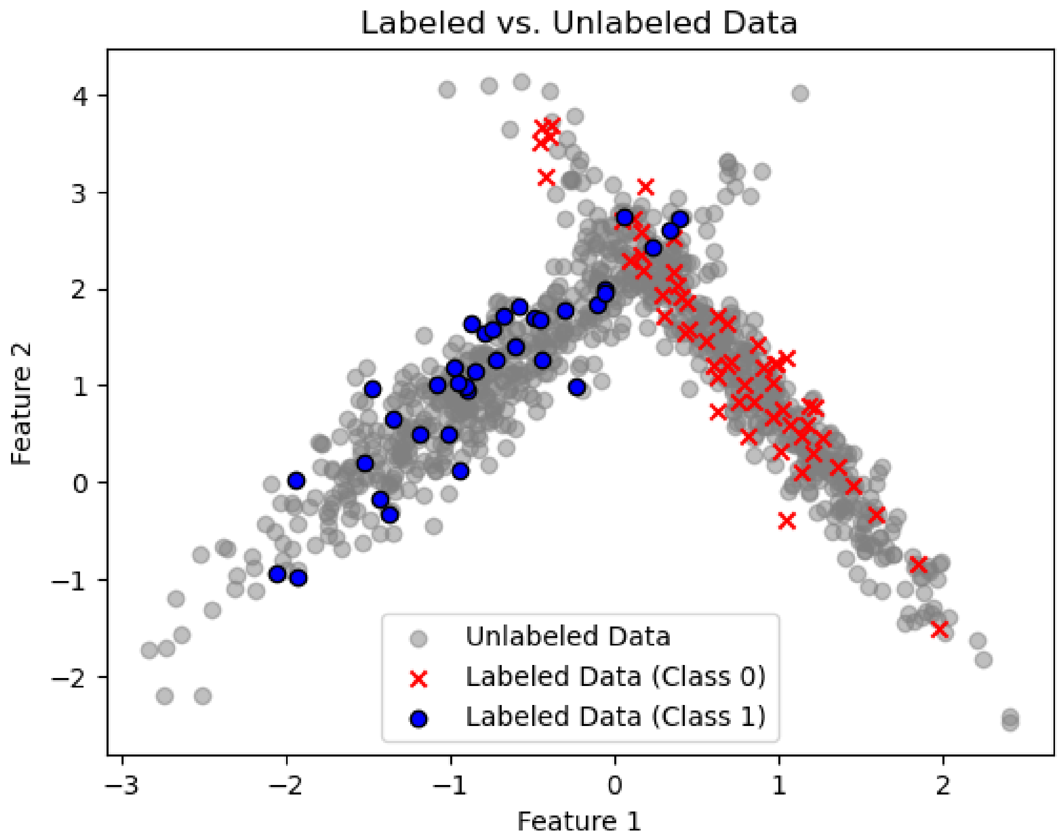
Characteristics:

- **Labeled Data**: Ground truth for training.
- **Unlabeled Data**: Will receive pseudo-labels during training.

Visualizing the Dataset

- **Labeled Data**: Points in red/blue (different classes).
- **Unlabeled Data**: Gray points.

Dataset Visualization



Key Observations from the Dataset

- **Labeled data:**
  - Sparse and covers limited regions.
  - Used to train the initial model.
- **Unlabeled data:**
  - Broader distribution.
  - Key source for pseudo-labeling.
- Visual shows the potential for improvement with pseudo-labeling.

Initial Model Training

Objective:

- Train a model using only labeled data.
- Use this model to generate predictions for the unlabeled data.

Initial Model Training

Process:

- **Train Model**:
  - Use labeled data to create an initial classifier.
- **Predict on Unlabeled Data**:
  - Use the model to predict labels for the unlabeled samples.
- **Limitations**:
  - Sparse labeled data leads to less generalizable decision boundaries.

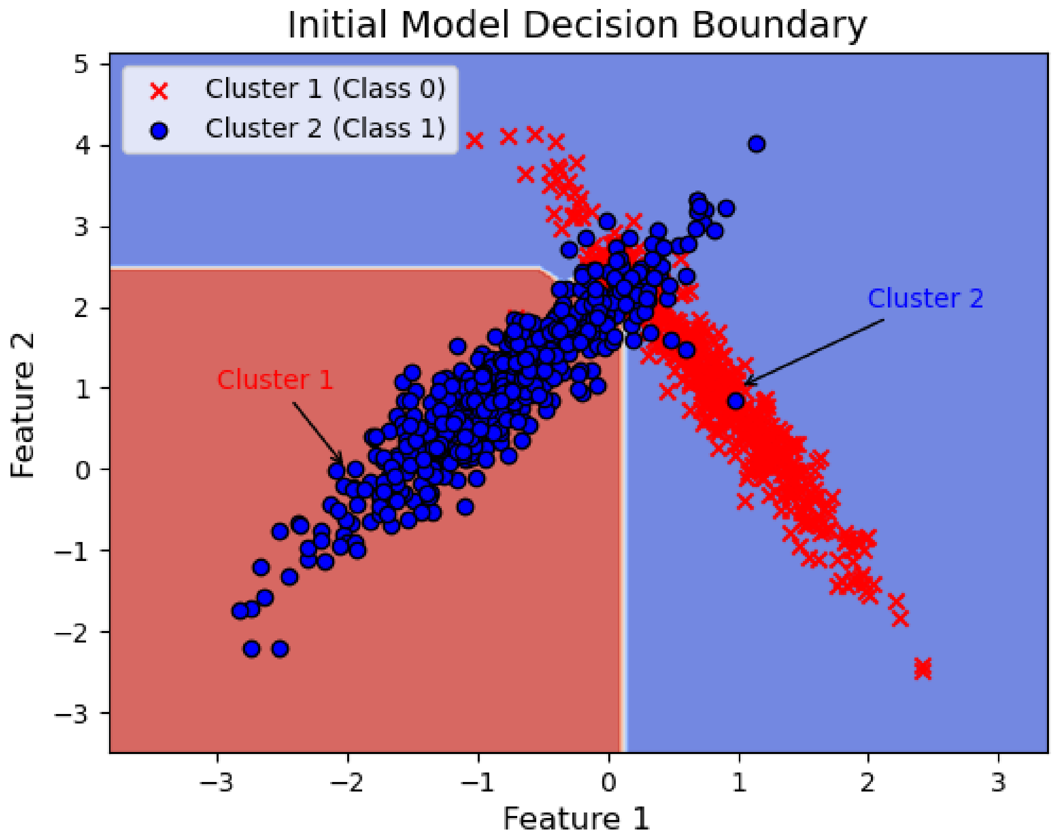
Initial Model Training

Visualizing the Decision Boundary:

- Initial decision boundaries are based on limited labeled data.
- This often results in inaccurate or uncertain regions.

Observations:

- The initial decision boundary is overly simplistic.
- Limited labeled data causes poor generalization to the unlabeled data.



Introducing Confidence Thresholding

Objective:

- Address the issue of noisy pseudo-labels.
- Use confidence thresholding to filter out low-confidence predictions.

What is Confidence Thresholding?

- A method to ensure only reliable pseudo-labels are used.
- **Process**:
    - The model assigns confidence scores to its predictions.
    - Predictions with confidence scores above a defined threshold are retained.
    - Low-confidence predictions are discarded.

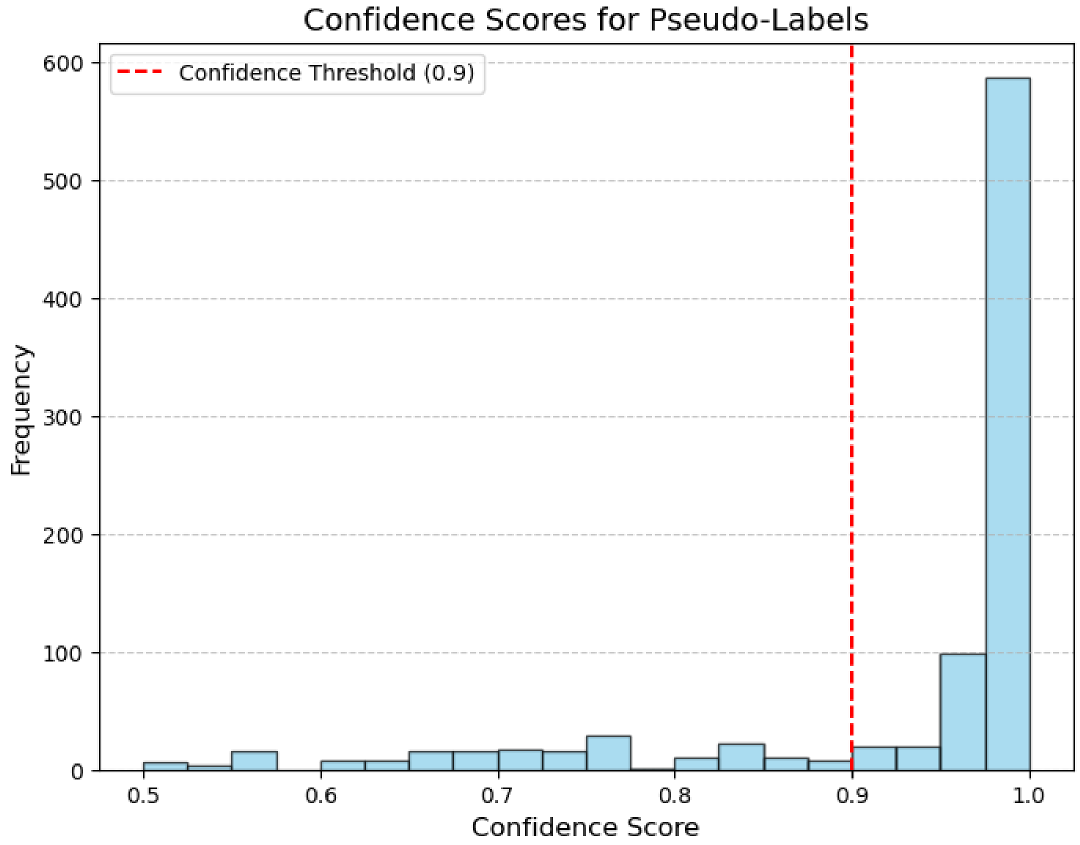
Introducing Confidence Thresholding

Benefits:

- Reduces the impact of noisy labels.
- Improves the quality of retraining data.

Visualizing Confidence Scores:

- A histogram of confidence scores for unlabeled data shows the distribution.
- The threshold (e.g., 0.9) separates high-confidence predictions.



Retraining the Model

Objective:

- Combine high-confidence pseudo-labeled data with labeled data.
- Retrain the model to improve decision boundaries.

Process:

- **Select High-Confidence Data**:
  - Filter pseudo-labels based on confidence scores.
  - Merge them with labeled data.
- **Retrain the Model**:
  - Use the combined dataset to refine the decision boundary.

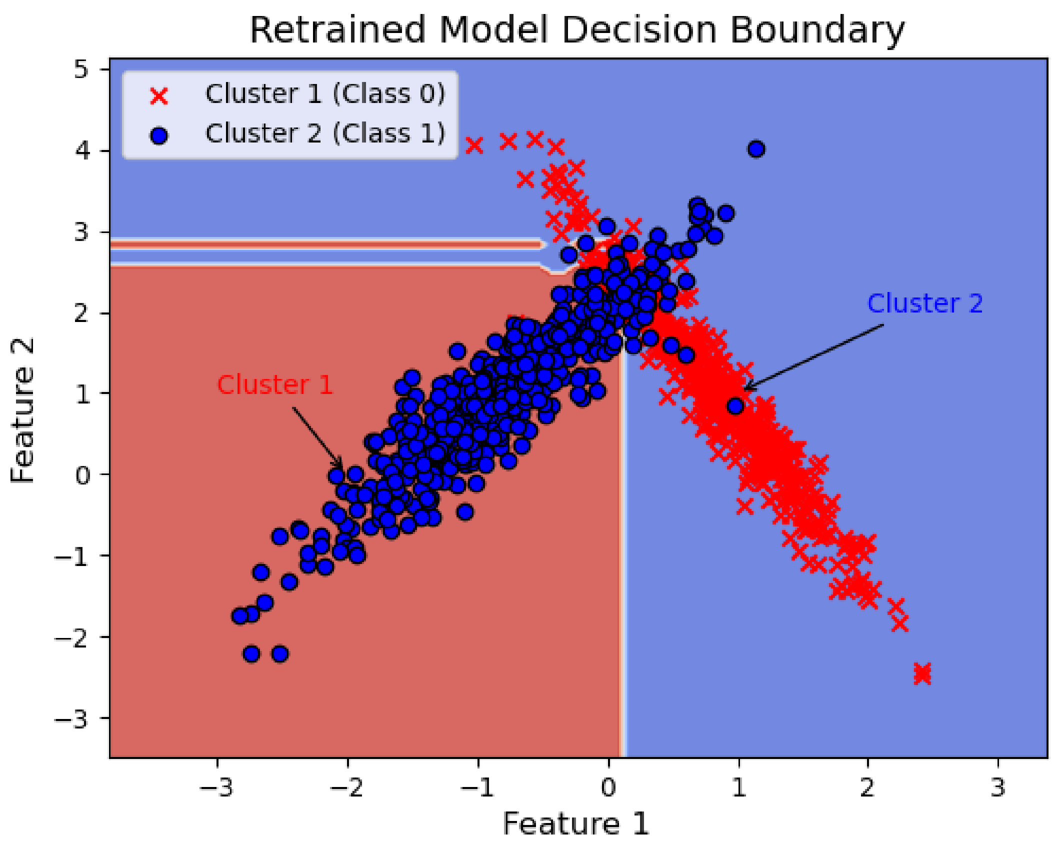
Retraining the Model

Visualizing the Updated Decision Boundary:

- Retraining adjusts the decision boundary to better fit the data.
- High-confidence pseudo-labeled data helps expand the boundary into previously uncertain regions.

Observations:

- The updated decision boundary is more accurate.
- Incorporating high-confidence pseudo-labeled data reduces uncertainty in classification regions.



Conclusion

Key Insights:

- **Self-Training**:
  - A powerful semi-supervised learning technique.
  - Iteratively improves performance using pseudo-labeled data.
- **Confidence Thresholding**:
  - Ensures the reliability of pseudo-labels.
  - Balances between quantity (lower threshold) and quality (higher threshold).
- **Practical Implications**:
  - Works best with datasets where labeled data is scarce but informative.

Conclusion

Recommendations:

- Experiment with different confidence thresholds.
- Monitor the distribution of pseudo-labels to avoid imbalance.
- Combine confidence thresholding with other semi-supervised methods for enhanced performance.

References

- Lee, D.H. (2013). Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method. *arXiv preprint arXiv:1301.0796*. Available at: https://arxiv.org/abs/1301.0796.
- Zhu, X. (2008). Semi-Supervised Learning Literature Survey. *University of Wisconsin-Madison*. Available at: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf.
- Scikit-learn Documentation (2023). *SelfTrainingClassifier*. Available at: https://scikit-learn.org/1.5/modules/generated/sklearn.semi_supervised.SelfTrainingClassifier.html