**SEMI-SUPERVISED SELF-TRAINING**

For the semi-supervised self-training, we used random forest classifier that we trained with the labelled data. After that we used that model to generate pseudo labels on the unlabeled data and retrained the model for these that got a higher confidence level than threshold.

Here are the results we got from it

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy | 0.95 | 0.92 | 0.90 | 0.89 | 0.88 |
| F1-score | 0.74 | 0.53 | 0.37 | 0.19 | 0.02 |

A graph of a curve

Description automatically generated

ST 50% unlabeled data

A graph of a curve

Description automatically generated

ST 75% unlabeled data

A graph with a line

Description automatically generated

ST 90% unlabeled data

A graph with a line and a blue line

Description automatically generated with medium confidence

ST 95% unlabeled data

A graph with blue and orange lines

Description automatically generated

ST 99% unlabeled data

**SEMI-SUPERVISED CO-TRAINING**

For the semi-supervised co-training, we used random forest and gradient boost as base classifiers that we trained with the labelled data. After that we used these models to generate pseudo labels on the unlabeled data and retrained each model 10 times for the 10 labels that got the highest confidence level by the other classifier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy-clf1 | 0.87 | 0.87 | 0.87 | 0.87 | 0.92 |
| F1-score-clf1 | 0.23 | 0.17 | 0.14 | 0.12 | 0.0 |
| Accuracy-clf2 | 0.91 | 0.90 | 0.88 | 0.88 | 0.90 |
| F1-score-clf2 | 0.53 | 0.49 | 0.5 | 0.34 | 0.32 |

A graph of a curve

Description automatically generated with medium confidence

CT CLF 1 50% unlabeled data

A graph with a line

Description automatically generated

CT CLF 1 75% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 1 90% unlabeled data

A graph with a line

Description automatically generated

CT CLF 1 95% unlabeled data

A graph with a line and a blue dotted line

Description automatically generated

CT CLF 1 99% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 50% unlabeled data

A graph of a curve

Description automatically generated with medium confidence

CT CLF 2 75% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 90% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 95% unlabeled data

A graph of a positive rate

Description automatically generated with medium confidence

CT CLF 2 99% unlabeled data

In the exploration of semi-supervised learning methodologies, the comparative effectiveness of self-training and co-training approaches is assessed. The analysis centers around their performance at incrementally higher proportions of unlabeled data, specifically 50%, 75%, 90%, 95%, and 99%. The self-training model employs a Random Forest classifier, whereas the co-training model leverages both Random Forest and XGBoost classifiers, harnessing their distinctive learning capabilities.

Self-Training Model Evaluation:

The self-training model demonstrated high resilience across the spectrum of unlabeled data proportions. The accuracy remained notably high, with the lowest recorded at 94.77% with 50% unlabeled data and the highest at 96.41% with 90% unlabeled data, suggesting the robustness of the Random Forest algorithm in self-training contexts. The F1-score, a more balanced measure considering both precision and recall, also showed a high level of performance, though with greater variation: a peak of 74.48% at 50% unlabeled data and a notable decrease to 36.77% at 90% unlabeled data, reflecting the increased difficulty of maintaining performance as the proportion of unlabeled data rises. The Receiver Operating Characteristic (ROC) curves, paired with Area Under the Curve (AUC) metrics, affirm the classifier's proficiency. Despite a slight decline in AUC with increased unlabeled data, the curves stayed significantly above the baseline, indicating an effective model.

Co-Training Model Evaluation:

In the co-training model, the XGBoost classifier consistently outperformed the Random Forest classifier in terms of accuracy, with the former attaining a peak accuracy of 95.87% at 75% unlabeled data compared to the latter's peak of 87.90% under similar conditions. The F1-scores for XGBoost ranged from a high of 53.89% at 75% unlabeled data to a low of 43.14% at 95% unlabeled data. The Random Forest classifier experienced a decrease in F1-score from 23.16% to a mere 8.18% as the unlabeled data proportion increased from 50% to 95%. ROC curves for the co-training method indicated a disparity in classifier performance, with the XGBoost classifier achieving higher AUCs, thus underscoring its superiority in discriminating between the classes across varied thresholds.

Comparative Insights:

The comparative analysis elucidates several key findings:

* Algorithmic Efficacy:

Within the co-training paradigm, XGBoost emerges as the more effective algorithm, demonstrating superior accuracy and F1-scores relative to the Random Forest classifier.

* Method Complexity:

The complex nature of the co-training method did not unequivocally translate to superior performance compared to the self-training approach, indicating that a single well-tuned classifier can sometimes match or exceed the performance of two combined classifiers.

* Metric Interpretation:

Accuracy alone is insufficient for thorough model evaluation. The F1-score and ROC analysis provide a nuanced interpretation of performance, particularly in scenarios with significant class imbalances.

* Data Proportionality Impacts:

As the ratio of unlabeled to labeled data increases, there is a discernible impact on model performance metrics, signaling the importance of the quantity and quality of labeled data in semi-supervised learning frameworks.

Conclusive Remarks:

The analysis affirms the potential of self-training with Random Forest as a compelling semi-supervised learning technique, attributable to its simplicity and consistent high performance. In contrast, while co-training with XGBoost and Random Forest offers advantages, particularly with XGBoost in specific scenarios, it does not consistently outperform the self-training approach.