School of Electrical Engineering and Computer Science

CSI5155 – Machine Learning

Project Fall 2024

Total: 100 marks

Topic: Semi-supervised learning and label scarcity.

Instruction:

1. Complete this project on your own or in a group of two (2) students.
2. Submit your project using BrightSpace before the due date.
3. We cannot accept late submissions.
4. Each group member must individually submit the project.
5. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that if you choose to use GitHub, the date and time of the last change to your repository should be **before** the deadline.
6. All students are required to demonstrate their projects during a timeslot that the teaching assistant will schedule.

This project aims to study the interplay between semi-supervised learning (SSL) and label scarcity.

Datasets and Tasks:

Use the Magic Mushrooms dataset from assignments 1 and 2 [1].

A: Semi-supervised learning [50 marks]

Semi-supervised learning addresses the scenario where most class labels are unknown and is a common technique used in real-world applications. This approach combines a small amount of labelled data with a large amount of unlabeled data during training. Semi-supervised learning is based on the observation that unlabeled data can produce considerable performance improvements when used with a small amount of labelled data.

1. Supervised learning (5 marks)

As a first step, use the Gradient Boosting algorithm to construct a supervised learning model that will form the baseline against which the semi-supervised learning models will be compared. You should aim to obtain the highest overall accuracy.

Follow the holdout method, with a split of 67% training set and 33% test set, as is standard practice in semi-supervised learning experiments.

- Use the original dataset without any form of sampling.
- Be sure to stratify your training and test set split to maintain the proportion of majority and minority class instances in both the training and test sets.
- If you have not done so during assignment 1, conduct experiments to determine the most appropriate number of estimators and learning rate. For instance, you could use Grid Search, Random Search, or Bayesian optimization with libraries such as Scikit-Optimize or Optuna.
-

2. Semi-supervised learning (40 marks)

In this step, you are required to implement four (4) different inductive semi-supervised approaches, one algorithm each from the following list [2-4]:

1. a self-training algorithm using labelled data to iteratively label unlabeled instances.
2. a co-training algorithm where two classifiers iteratively label each other's unlabeled instances.
3. a semi-supervised ensemble such as the SemiBoost algorithm [4].
4. an approach that employs unsupervised pretraining or an intrinsically semi-supervised learning method other than the semi-supervised ensemble you selected in 3.

Please refer to Van Engelen and Hoos's survey [2] for a detailed discussion on semi-supervised learning. The survey also includes descriptions of and references to many algorithms within the various categories.

B: Evaluation of results [25 marks]

An essential step in semi-supervised learning is determining the level of labels the algorithm needs to perform well. The literature suggests that, for most domains, the number of labelled data should be between 10% and 20%. Once you have created the models, you should thus test various levels of unlabelled data. Specifically, you should follow the following steps.

1. Split the Dataset into Training and Testing Sets

- Use exactly the same dataset split as when constructing a model using the Gradient Boosting algorithm in Step A(1).
- The test set remains fully labelled and is used only for final evaluation. This fully labelled test set is a consistent benchmark for evaluating the model's performance, allowing you to measure metrics like accuracy, precision, recall, and F1 score without the uncertainty introduced by unlabeled data.
- By keeping the test set fully labelled, you can directly assess how well the model generalizes and how effectively it has learned from the combination of labelled and unlabeled training data. This setup is essential to reliably compare performance across different labeled-unlabeled splits in the training set.
- The training set will have labels incrementally removed to simulate varying levels of labelled and unlabeled data.

2. Define Labeled and Unlabeled Subsets

- From the training set, randomly sample a small portion to retain as labelled data and designate the rest as unlabeled. Initially, start with 10% labelled and 90% unlabelled.
- We will incrementally decrease the number of unlabelled instances. We will repeat the experiment multiple times, i.e., starting with 10% labelled, then 20%, 30%, 40%, and 50%.
- This step-by-step increase allows analysis of how the algorithm's performance changes as more unlabeled data is included.

3. Train the Semi-Supervised Model

- For each labelled-unlabeled split in Step 2, train the semi-supervised learning algorithms using the labelled and unlabeled data in the training set.
- Some algorithms may require a pre-training phase on the labelled data and an iterative self-training or label propagation phase to leverage the unlabeled data.

4. Evaluate the Model on the Test Set

- After training on each labelled-unlabeled split, evaluate the models' performances using the fully labelled test set.

- You should report accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).
- Be sure to compare the results of semi-supervised models to the Gradient Boosting supervised baseline, i.e., trained only on the labelled portion of the training set without any unlabeled data. This helps illustrate how performance is affected by the presence of unlabeled data and can demonstrate the effectiveness of semi-supervised approaches.

C: Source code, final report, and project demonstration [25 marks]

Submit your source code (or a link to a GitHub repository).

Submit a PDF file containing the tables and figures of the evaluation results in B, together with a 500-word report that discusses the results you obtained and the lessons you learned.

References / Link to resources:

[1]. Drug Consumption Dataset for experimentation on predictive modelling: https://link.springer.com/book/10.1007/978-3-030-10442-9 or https://arxiv.org/abs/1506.06297 .

[2]. A survey on semi-supervised learning: J.E. van Engelen, J.E. and H.H. Hoos, Machine Learning Journal, 109, 373–440, https://doi.org/10.1007/s10994-019-05855-6, 2020.

[3]. Semi-supervised techniques in Scikit Learn: https://scikit-learn.org/stable/modules/semi_supervised.html

[4]. SemiBoost: Boosting for Semi-Supervised Learning: P. K. Mallapragada, R. Jin, A. K. Jain and Y. Liu, https://ieeexplore.ieee.org/abstract/document/4633363