Task 2 Report

Sam Reswinraj Abraham (s4248325)
Faculty of Science and Engineering
University of Groningen)
Groningen, The Netherlands
a.sam.reswin.raj@student.rug.nl

Maniraj Sai Adapa (s4574842)
Faculty of Science and Engineering
University of Groningen)
Groningen, The Netherlands
m.s.adapa@student.rug.nl

Thijs Visee (s2982129)
Faculty of Science and Engineering
University of Groningen)
Groningen, The Netherlands
t.p.visee@student.rug.nl

Niels Rocholl (s3501108)
Faculty of Science and Engineering
University of Groningen)
Groningen, The Netherlands
n.m.m.rocholl@student.rug.nl

A. Introduction

This report covers task 2 of assignment 2 for the pattern recognition 2021-2022 course. This task concerns applying both a baseline supervised and a semi-supervised learning method on a sparsely annotated data-set, comparing their performances afterwards.

B. Data

The dataset is a csv-file containing 284,807 datapoints for credit card transactions, containing the time and amount of the transaction, a 28-dimensional vector representing the encoded credit card values and a boolean class label, '1' indicating a fraudulent transaction, '0' indicating a normal, legal transaction.
The data is split into three parts, randomly, whilst keeping the ratio between fraudulent and normal transactions constant between all parts. This results in the sets `X_train_labeled` (24%), `X_train_unlabeled` (56%) and `X_test` (20%). After creating this division, a number of random non-fraudulent datapoints in `X_train_labeled` and `X_test` are removed, after which both classes have an equal amount of samples in those sets.

C. Performance Metric

The trained models are tested with `X_test`, the given prediction and the ground truth labels are used to calculate the F2-score, for which the weight of the sensitivity is equal to the weight of the specificity.

D. Supervised Learning

The used supervised learning algorithm is a logistic regression model, imported from the python3 sklearn library. The training input are the 28-dimensional vectors, the training output are the corresponding labels. Class weights or regularization techniques are not used. The performance of the trained logistic regression model is measured using -C.

E. Semi-supervised Learning

The algorithm used for the semi-supervised learning is a label propagation based on *K-Nearest Neighbours (KNN)*, imported from the sklearn library. The three closest neighbours are considered in the aforementioned KNN application. After performing the label propagation on `X_train_unlabeled` (which is now labeled), the non-fraudulent transactions in `X_train_unlabeled` are undersampled obtain an equal representation of both classes in the set. By concatenating the newly labeled `X_train_unlabeled` with `X_train_labeled`, a new training set is created to train the logistic regression model.
The performance of the trained label propagation and logistic regression models are computed using -C.

F. Results

All methods are performed 100 times, yielding mean F2-scores of 0.927, 0.932 and 0.938 for the baseline, label-propagation and semi-supervised methods respectively (I), with standard deviations of 0.020 (baseline), 0.016 (label-propagation) and 0.018 (semi-supervised).

F2-Score	Baseline	Label-Propagation	Semi-Supervised
Mean	0.927	0.932	0.938
Standard Deviation	0.020	0.016	0.018

TABLE I: The F2-Score means and standard deviations of all three methods over 100 runs.

The performances of the three methods are plotted in 1. The index of an iteration is irrelevant to the performance, therefore the baseline results have been sorted to follow an ascending order, with the label-propagation and semi-supervised method results following the same permutations, in order to create a figure that is more intuitive to interpret.

G. Discussion

The representation of both classes in the dataset is highly unbalanced. Balancing the test set, as hinted upon in the

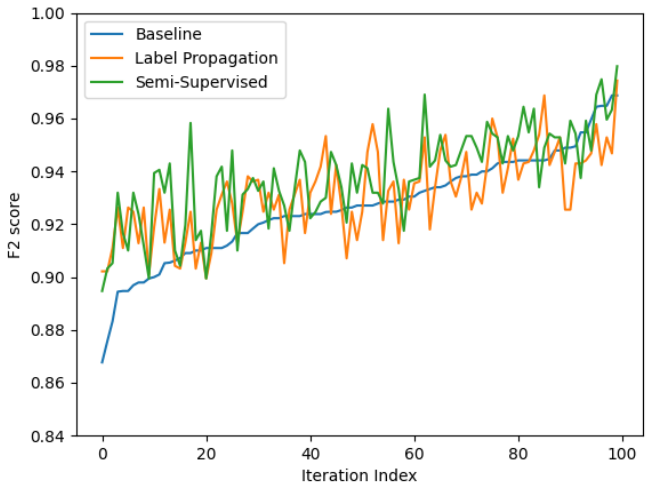

Fig. 1: F2-score for Baseline, Label Propagation and the full Semi-Supervised approaches. Iterations are sorted by baseline performance

assignment, creates an unrealistic assumption of the model's practical risk. However, on the grounds of the general, though not consistent, better performance of the semi-supervised method it can be concluded that it is beneficial to use semi-supervised learning in similar scenarios. The extra computational expense is limited, but weighted against the improved performance the semi-supervised method is our preferred choice.

Author contributions

All work was divided equally.