Assignment 2, task 2: semi-supervised learning

Tim Eckhart (s3952509)[1], Kyriakos Antoniou (s5715881)[1], Robert Power (s5332419)[1], and Abi Raveenthiran (s4010132)[1]

Group [06]
Pattern Recognition (WMAI021-05) 2023-2024.1A
[1] Artificial Intelligence Department, University of Groningen
{*t.eckhardt,k.antoniou,r.power,a.raveenthiran*}@student.rug.nl

August 29, 2025

Answers to the questions (1-4)

This report focuses on a semi-supervised classification task. Specifically, on the Credit Card Fraud Detection dataset [2]. Semi-supervised learning classifiers [3] have a unique trait that traditional classifiers do not. That is, that semi-supervised classifiers can be trained on datasets where a significant portion of the data is unlabeled. In the case of fraud detection, this can be very useful as annotating fraudulent transactions requires a lot of resources (i.e., time, money, effort). The objective of this task is to explore if using a semi-supervised approach is useful when dealing with fraud detection with he given dataset.

The dataset, D, consists of encoded transactional data, X, with two classes, y, of non-fraudulent and fraudulent transactions. The dataset consists of 284807 transactions with 492 of them being fraudulent. The class distribution of non-fraudulent to fraudulent transactions is $0.99827 : 0.00173$, clearly indicating a highly imbalanced dataset. Preprocessing of the dataset was done by dropping the *Time* and *Amount* coloumns. The input data consists of 28 encoded credit card data features. The dataset was split randomly into two subsets: a train set, T, with 80% of D, and a test set, G, with the remaining 20% of D. The class distribution of original dataset was preserved during the train-test split using Stratified k-fold [1]. The train set was further split into two subsets, a labeled train set with 30% of T, and an unlabeled train set with the remaining 70% of T. Once again, the class distribution of the train set was preserved using Stratified k-fold. The class labels associated with the unlabeled train set were dropped for the subset to be indicative of unlabeled data.

To tackle the objective of this task, three experimental cases were implemented:

- **Case 1**: Training a baseline model solely on the labeled train set. A k-Nearest Neighbours (k-NN) classifier [1] was used for the baseline model. Grid search [1] was used to find optimal k in the arbitrarily chosen range $k \in [1, 10]$.
- **Case 2:** Training a semi-supervised model on the combined set of the labeled train set and the unlabeled train set. A Label Propagation classifier, based on k-NN with the same optimal k found in the grid search, was used for the semi-supervised model.

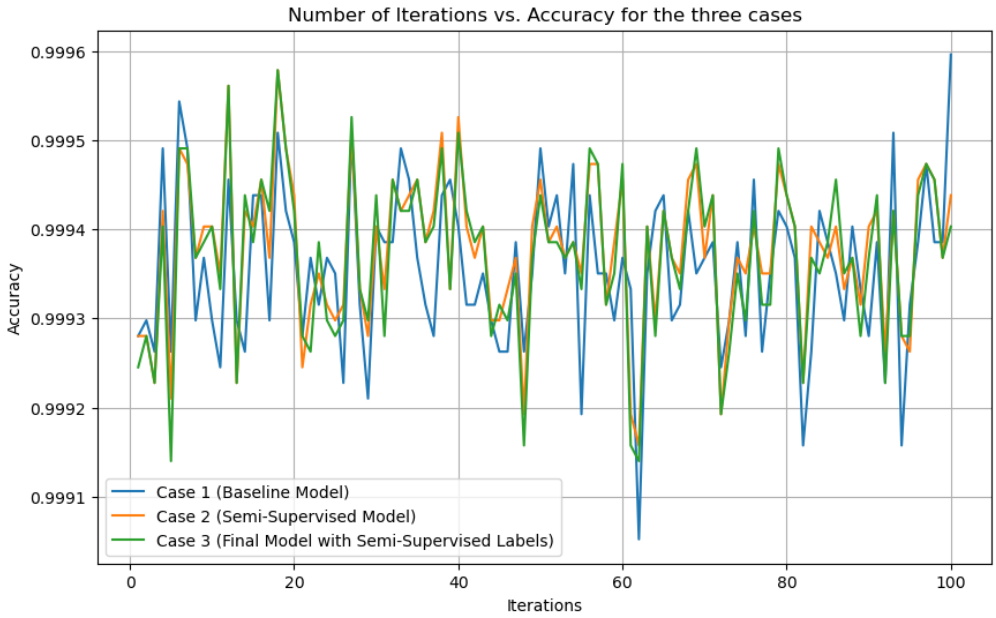

Figure 1: Simulation Accuracy of the three experimental cases

Experimental Case	Mean Accuracy	Mean F1 Score
Case 1	0.99936	0.79607
Case 2	**0.99938**	**0.80278**
Case 3	0.99370	0.80026

Table 1: Simulation Mean Performance of the three experimental cases

- **Case 3:** Training a final model on the combined set of the labeled train set and the unlabeled train set, but, using class labels produced by the semi-supervised model through transduction [1]. A k-NN classifier was used for the final model.

The experimental setup involved simulating the three cases 100 times to obtain accurate and robust results. The simulation is stochastic due to the random splitting of the data. Two evaluation metrics were employed, namely *accuracy* and *F1 score*. The results of the simulation in the context of the aforementioned evaluation metrics are presented in Figure 1, Figure 2, and Table 1.

Discussions and/or conclusions

The accuracy results depicted in Figure 1 and Table 1 are very high. These results seem misleading due to the highly imbalanced nature of dataset. In the context of this task, accuracy is not an appropriate evaluation metric to determine whether the classifiers are effectively predicting fraudulent transactions. The F1 score on the other hand provides a more accurate representation of the effectiveness of the experimental cases. The F1 scores depicted in 2 and Table 1 show that case 2 is preferable. This would need to be confirmed through a formal hypothesis test at a specified confidence level for credibility. However, in terms of the previously stated objective of this task, the

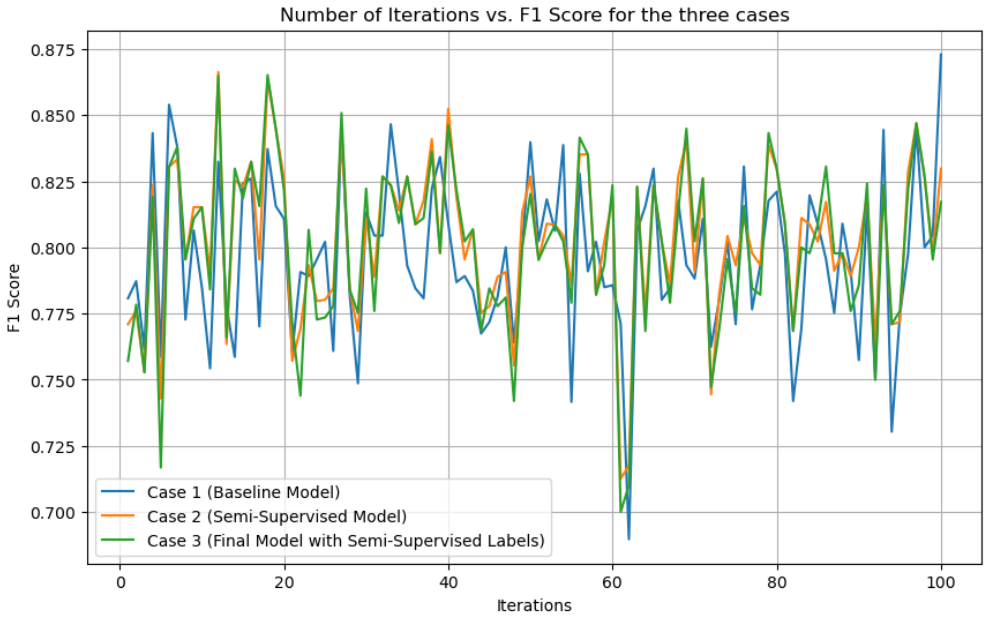

Figure 2: Simulation F1 Score of the three experimental cases

given dataset is already completely labeled, so a semi-supervised approach is not necessary. The labels obtained from Case 2 and used in Case 3 provide a very minor improvement on F1 score at the expense of a significant increase in computational cost. Thus, reinforcing the notion that a semi-supervised approach is not helpful in this type of data to predict the frauds of credit cards.

Individual contributions

This task was done by Robert Power (s5332419) and reviewed by the other members of Group [06].

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[2] Machine Learning Group Université Libre de Bruxelles Worldine. Credit card fraud detection. `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`.

[3] XiaojinZhu. Semi-supervised learning literature survey. Technical Report 1530, Computer-Sciences, University of Wisconsin-Madison, 2008.