Detecting frauds through statistics and machine learning: an overview of supervised and unsupervised algorithms

Dipartimento di Metodi e Modelli
per l'Economia il Territorio e la Finanza
MEMOTEF

Sapienza
Università di Roma

Domenico Vitale - domenico.vitale@uniroma1.it

27 June 2024

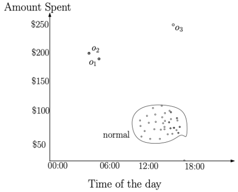
Fraud detection as anomaly detection

- Frauds can be cast as deviations from normal transaction data and can be addressed using anomaly detection procedures
- Anomaly detection is a broad field that addresses the problem of identifying instances of data or events that do not conform to expected behaviour
- An anomaly is an observation which deviates so much from the other observations as to arouse suspicious that it was generated by a different mechanism (Hawkins, 1980)

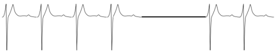
Types of anomalies

Anomalies can be classified into three different categories (Chandola, Banerjee, & Kumar, 2009):

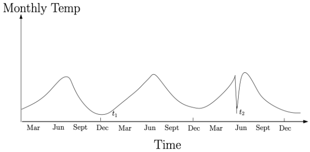
- Point anomalies: an individual instance is anomalous with respect to the data



- Collective anomalies: a collection of related data instances is anomalous



- Contextual anomalies: an individual data instance is anomalous within a context



Challenges in Fraud Detection

- It is very difficult to define a normal region or boundary to encompass all possibilities of normal behaviour, and usually, the boundary between normal and anomalous behaviour lacks precision
- Anomalies that arise due to malicious activity are often changing and adapting (concept drift), driven by adversaries of the anomaly detection system and their attempts to disguise anomalous events as normal, ultimately increasing the difficulty of detection
- The notion of an anomaly varies from different domains and applications. For this reason, applying a technique that is developed for one domain may not be as straightforward to implement in another
- Lack of labeled data for training and validation of models due to several reasons (eg sensitive data or costs)
- Anomalous instances are also rare in occurrence, contrasted by normal instances. In such cases, standard classifier anomaly detection techniques tend to ignore the small classes due to being overwhelmed by the larger ones
- In low-dimensional spaces, anomalies often display prominent abnormal features or characteristics. However, they become hidden and indiscernible in high-dimensional spaces (curse of dimensionality)

Performance measures limitations

The performance evaluation of anomaly detection algorithms relies on metrics like:

- Precision - % of detected anomalies which are true anomalies
- Recall - % of actual anomalies successfully detected
- F1 score - Balance of precision and recall
- AUROC

However, these metrics require labeled data, thus are useful only for supervised anomaly detection algorithms

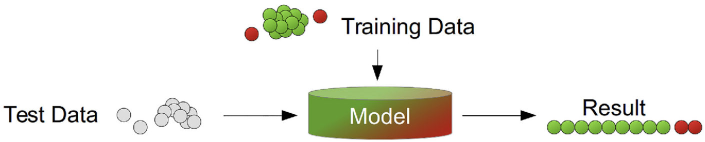
Supervised, Semi-Supervised and Unsupervised Anomaly Detection

Which is the best performing anomaly detection algorithm for fraud detection?

In the following we report the main results highlighted in a recent paper by
Hilal et al (2022) Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances, Expert Systems With Applications 193 116429
https://doi.org/10.1016/j.eswa.2021.116429(217citations)

Supervised, Semi-Supervised and Unsupervised Anomaly Detection

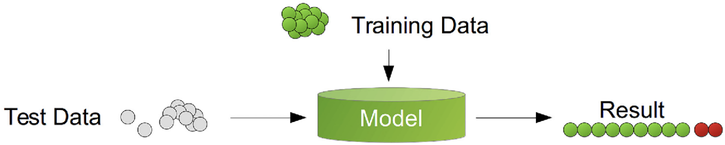
- Supervised anomaly detection models are designed to detect anomalies in dataset with labeled examples of anomalies and normal data points



(a) Supervised anomaly detection

Supervised, Semi-Supervised and Unsupervised Anomaly Detection

- Semi-supervised anomaly detection models assume that the only instances in the data set that are labeled are the ones belonging to the normal class. A model is constructed only for the normal class and not the anomalous class. The test set of the data is then compared against the model to identify anomalous instances



(b) Semi-supervised anomaly detection

Supervised, Semi-Supervised and Unsupervised Anomaly Detection

- Unsupervised anomaly detection models do not require any labels in the data set. An implicit assumption is made by unsupervised methods that anomalous events are far less frequent than normal events in the test set of the data

(c) Unsupervised anomaly detection

Supervised methods for fraud detection

Several supervised anomaly detection algorithms have been applied for fraud detection (Waleed et al 2022[1]), mainly based on

- Support Vector Machine (SVM)
- Neural Networks (NN)
- Convolutional Neural Networks (CNN)
- Long Short-Term Memory Networks

[1] Hilal W, Gadsden SA, Yawney J (2022) Financial fraud: a review of anomaly detection techniques and recent advances. Expert Systems with Applications, doi: 10.1016/j.eswa.2021.116429

Supervised methods for fraud detection: literature review

Such methods where adapted to different types of fraud and performed on labeled dataset

Table 5
Summary of published literature on SVM-based fraud detection.

Year	Reference	Type of fraud	Method	Comments
2011	(Sahin & Duman, 2011)	Credit card	SVM	SVM with stratified sampling overfit the data and outperformed by DT.
2011	(Bhattacharyya et al., 2011)	Credit card	SVM	LR outperforms SVM. As fraud rate decreases, results become comparable.
2011	(Lu & Ju, 2011)	Credit card	ICW-SVM	ICW-SVM superior to SVM and DT, and computationally more efficient.
2013	(Hejazi & Singh, 2013)	Credit card	OCSVM	One-class SVM outperforms SVM in imbalanced data sets.
2020	(Rtayli & Enneya, 2020)	Credit card	RF and SVM	RF-SVM ensemble accuracy comparable to but less than LOF-IF, however, demonstrated the highest AUC.
2012	(Tao, Zhixin, & Xiaodong, 2012)	Auto-insurance	DFSVM	DFSVM outperforms vanilla SVM in terms of F-score, recall and precision.
2015	(Sundarkumar & Ravi, 2015)	Auto-insurance	*k*RNN-OCSVM [a]	Notable increase in AUC and recall for SVM model, with loss in precision.
2016	(Sundarkumar, Ravi, & Siddeshwar, 2015)	Auto-insurance	OCSVM [a]	*k*RNN identified to slightly limit overall performance, therefore eliminated.

[a] The denoted methods proposed are SVM-based undersampling techniques augmented with a fraud detection system rather than actual classifiers.

Table 7
Summary of Published Literature on NN-Based Credit Card Fraud Detection.

Year	Reference	Type of Fraud	Method	Method proposed
1993	Maes *et al.* (Maes & Tuyls, 2002)	Credit card	MLP	MLP trained on preprocessed dataset produced good results but was outperformed by a Bayesian network. Adaptive learning rate proved to be beneficial.
1994	Ghosh and Reilly (Ghosh & Reilly, 1994)	Credit card	MLP	MLP resulted in 20 to 40 percent decrease in economic losses.
1997	Aleskerov *et al.* (Aleskerov et al., 1997)	Credit card	MLP	Using momentum during training improved MLP performance, detecting 85 percent of fraud cases.
2011	Patidar and Sharma (Patidar & Sharma, 2011)	Credit card	GA-designed MLP	Theoretical research proposing GA to address lack of guidelines for selecting number and size of hidden layers to use in a network.
2014	Khan *et al.* (Khan, Akhtar, & Qureshi, 2014)	Credit card	SA-trained MLP	Model achieves high detection rate at cost of increased false positives and is computationally expensive.
2015	Behera and Panigrahi (Behera & Panigrahi, 2015)	Credit card	FCM-MLP	FCM for sample filtering, then MLP trained with SCG to classify suspicious achieved 94 percent accuracy, only 6 percent false alarm rate.
2018	Wang *et al.* (Wang, et al., 2018)	Credit card	WOA-trained MLP	Model generalizes well and addresses problems of NNs overfitting with F-score of 98.4 percent.
2018	Gómez *et al.* (Gómez et al., 2018)	Credit card	MLP ensemble	Ensemble of MLP filters reduce effects of imbalanced data, improve classification performance of classifier.

Table 9
Summary of published literature on CNN and LSTM-based fraud detection.

Year	Reference	Type of fraud	Method	Comments
2016	Fu *et al.* (Fu, Cheng, Tu, & Zhang, 2016)	Credit Card	CNN	CNN achieved F-score of 0.33 and outperformed MLPs.
2017	Heryadi and Warnars (Heryadi & Warnars, 2017)	Credit Card	CNN-LSTM	CNN's short-term and LSTM's long-term abilities combined to capture temporal relations. Best AUC achieved of 77%.
2018	Zhang *et al.* (Zhang et al., 2018)	Credit Card	CNN	CNN achieved recall of 94% and precision of 91%, outperforming MLP, but was considerably slower in training.
2009	Wiese and Omlin (Wiese & Omlin, 2009)	Credit Card	LSTM	LSTM outperformed SVMs, as well as the MLP proposed by Maes *et al.* in (Maes & Tuyls, 2002).
2018	Jurgovsky *et al.* (Jurgovsky et al., 2018)	Credit Card	LSTM	LSTM with feature aggregation strategy performed similarly to RF but detected different fraud behaviours. Combination of models suggested.

Source: Hilal W, Gadsden SA, Yawney J (2022) Financial fraud: a review of anomaly detection techniques and recent advances. Expert Systems with Applications, doi: 10.1016/j.eswa.2021.116429

Unsupervised methods for fraud detection

Recent research adopted unsupervised and semi-supervised anomaly detection algorithms (Hilal et al 2022), and in particular

- Autoencoders (AE)
- Generative Adversarial Networks (GAN)

Table 11
Summary of published literature on autoencoder-based fraud detection.

Year	Reference	Type of fraud	Method	Comments
2017	Kazemi and Zarrabi (Kazemi & Zarrabi, 2017)	Credit card	AE	AE accuracy of 81.6% was outperformed by SOM accuracy of 82.4%
2018	Pumsirirat and Yan (Pumsirirat & Yan, 2018)	Credit card	AE	AE was superior to RBM, but performed poorly with small dataset size
2018	Sweers *et al.* (Sweers, Heskes, & Krijthe, 2018)	Credit card	VAE	AE with deeper architecture performed the best in terms of recall of 93.8% compared to VAE, both models had identical precision scores.
2018	Renström and Holmsten (Renström & Holmsten, 2018)	Credit card	Stacked AEs	Stacked AE and VAE models outperformed single AE model with a recall of 99% but had slightly lower precisions.
2019	Jiang *et al.* (Jiang, Zhang, & Zou, 2019)	Credit card	DAE-MLP[a]	DAE used to remove noise from input and use output to train MLP classifier. Outperformed MLP classifier trained on raw input.
2020	Misra *et al.* (Misra et al., 2020)	Credit card	AE-MLP[a]	AE using only the encoder for feature extraction, with the output used to train a classifier. AE-MLP classifier outperformed AE in (Pumsirirat & Yan, 2018)
2020	Tingfei *et al.* (Tingfei et al., 2020)	Credit card	VAE-MLP[b]	VAE to oversample minority class outperformed SMOTE and GAN oversampling when training MLP classifier.
2016	Paula *et al.* (Paula, Ladeira, Carvalho, & Marzagão, 2016)	Money laundering	AE	AE was able to detect fraudulent cases previously identified by domain experts

[a] The denoted methods are AE-based feature extraction or data preprocessing techniques implemented in conjunction with a classifier.
[b] The denoted methods are AE-based oversampling techniques augmenting generated data to a classifier's training set.

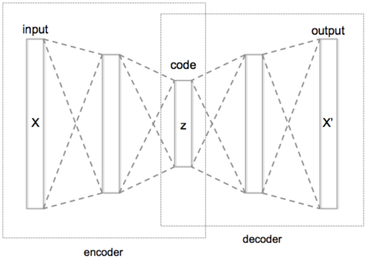
Table 12
Summary of Published Literature on GAN-Based Fraud Detection.

Year	Reference	Type of fraud	Method	Comments
2018	Chen *et al.* (Chen, Shen, & Ali, 2018)	Credit card	SAE-GAN	GAN trained on SAE-learned features from majority class has improved F-score and precision, but with a decrease in recall. SAE-GAN outperforms OCSVM and OCGP in terms of F-score.
2019	Tanaka and Aranha (Tanaka & Aranha, 2019)	Credit card	GAN-DT[a]	DT trained with minority class GAN-based oversampling had slightly higher precision, but lower recall than when using SMOTE or ADASYN.
2019	Fiore *et al.* (Fiore et al., 2019)	Credit card	GAN-MLP[a]	MLP trained with GAN-based oversampling had improved recall, and proposed model also outperformed SMOTE in terms of recall, but had slightly lower specificity.
2019	Ba (Ba, 2019)	Credit card	WCGAN-LR[a]	LR with WCGAN-based oversampling had most balanced performance with higher F-score and AUC than GAN, CGAN, SMOTE and ADAYSN. However, WCGAN's recall of 64.2% was significantly inferior to ADAYSN's at 90%.
2019	Zheng *et al.* (Zheng et al., 2019)	Credit card	AE-OCAN	Complementary GAN generator trained on AE-learned representations of genuine transactions; discriminator is proposed as OCAN. Proposed model performs better in F-score, precision and accuracy than OCSVM but outperformed in recall.
2020	Charitou *et al.* (Charitou et al., 2020)	Money laundering	SAE-GAN	SAE features extracted from entire train set, then used to train generator of GAN to produce complementary samples. Samples generated are augmented into training set, and discriminator is trained to classify samples. Proposed model outperformed LR, SVM, MLP and RF with either ADASYN or SMOTE in terms of F-score, accuracy and precision. RF with ADASYN, however, outperformed in terms of recall.

Autoencoders (AE) - Unsupervised

An autoencoder has a structure very similar to a feed-forward neural network, however, the primary difference when using in an unsupervised context is that the number of neurons in the output layer are equal to the number of inputs

Autoencoder based algorithms consist in two parts:
(1) an encoder function ($Z = f(X)$) that converts X inputs to Z codings and
(2) a decoder function ($X' = g(Z)$) that produces a reconstruction of the inputs (X')

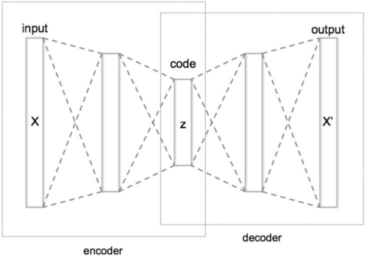


Autoencoders (AE)

To learn the neuron weights and, thus the codings, the autoencoder seeks to minimize some loss function (L), such as mean squared error (MSE), that penalizes X' (output) for being dissimilar from X (input): minimize $L = f(X, X')$

Since the loss function of an autoencoder measures the reconstruction error, we can extract this information to identify those observations that have larger error rates

Observations with large error rates have feature attributes that differ significantly from the other features, thus we might consider such features as anomalous, or outliers.



Generative Adversarial Networks (GAN)

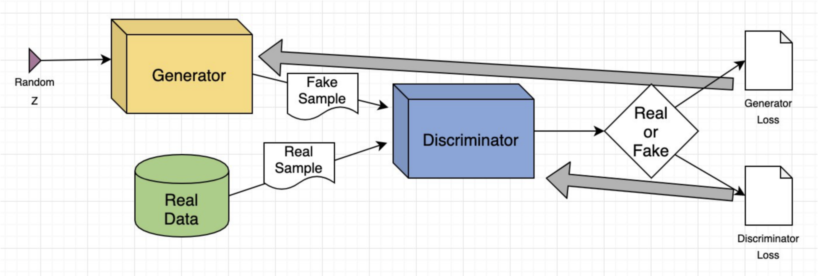
A GAN involves two deep neural networks: a generator and a discriminator network

Generator's Role: The generator aims to produce synthetic data that is so convincing that the discriminator cannot differentiate between real and generated data

Discriminator's Role: The discriminator is simultaneously trained to become more adept at distinguishing between real and generated data

The objective is for the generator to create data that is increasingly realistic, while the discriminator becomes more skilled at telling the difference. This adversarial process continues until the generator produces data that is essentially indistinguishable from real data

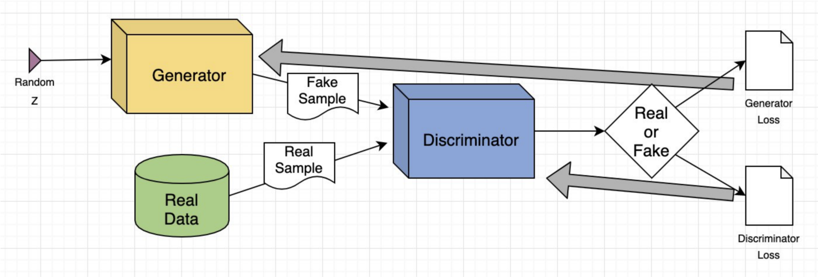
The equilibrium point, where the generator produces highly realistic data and the discriminator cannot reliably tell it apart from real data, represents the successful training of the GAN



Generative Adversarial Networks (GAN)

Anomaly detection through GAN can be performed in several ways

1. Selecting instances that are dissimilar to both the real and synthetic data
2. Selecting instances that the discriminator classifies more likely to be synthetic



Generative models (AEs and GANs) Pros and Cons

- Both GANs and AEs have proven to be superior in creating more realistic samples that capture a broader representation of data distributions for data augmentation than traditional oversampling approaches (eg MLP)
- These approaches have also proven to be preferable over those that involve undersampling the majority class, such as random undersampling, stratified sampling or even clustering algorithms dedicated to outlier detection and removal
- The limitations of deep learning models are that they require much more careful design and tuning compared to simpler models like SVM and RF, as they are rather sensitive to the choice of hyperparameters and the architecture structure

Final remarks

- The literature review showed that there is no single universally applicable anomaly detection technique or approach for all the different types of financial fraud
- From the surveyed literature, a clear shift in trend is apparent, with most of the recent research adopting unsupervised and semi-supervised models as opposed to supervised models
- An evident lack of publicly available datasets, labelled or not, was identified as a significant limitation in this field
- More importantly, the imbalanced nature of datasets due to the rare occurrence of fraudulent cases was emphasized as one of, if not the most critical considerations that must be factored in during the design stage of any fraud detection system or model
- Even when datasets are labelled, it is often the case that not all instances of fraud have been detected[2]

---

[2] Arezzo, MF, Guagnano G, Vitale D (2024) Estimating the size of undeclared work from partially misclassified survey data via the Expectation–Maximization algorithm. Journal of the Royal Statistical Society Series C: Applied Statistics 73.3 (2024): 816-834

Thanks for your attention!

Domenico Vitale - domenico.vitale@uniroma1.it

Dipartimento di Metodi e Modelli
per l'Economia il Territorio e la Finanza
MEMOTEF

Sapienza
Università di Roma