Banknote Authentication Design Considerations
Machine Learning for Security and Performance in Production Settings

Alice Chang, Karthik Umashankar, Yu Chen

June 5, 2018

Abstract

Banknote authentication is a high-value task within the financial services industry. In this paper, we perform a validation study and probe into potential vulnerabilities of online machine learning models utilized for this task. In particular, we utilize the work of "Banknote Authentication" by Gillich and Lohweg[1]. The authors utilized an industry-standard technique of filtering the banknote grayscale image for viable regions of interest, and then performing a novel Shift Invariant Wavelet Transform for the model's feature engineering. We first validate Gillich and Lohweg's findings of 100% accuracy on holdout datasets using a multitude of models (k-Nearest Neighbors, Support Vector Machine, etc.), and then identify critical thresholds where sample size begins to negative impact the potential performance of a machine learning model once deployed into production.

We conclude by investigating potential vulnerabilities an online machine learning system can encounter, namely the Frog Boiling Attack. We outline a potential implementation of this type of penetration attack and highlight various potential countermeasures to mitigate its negative impact on a machine learning system in production.

1 Introduction

The Intaglio technique used within banknote printing allows for an unmatched sharpness in contrast through a feelable relief[2] .This technique forms a vital component of modern-day banknote authenticity verification. In the initial phase of authentication, **image digitization** occurs where the banknote's relief is captured through industry-grade cameras to produce an image grayscale that can be then undergo a feature transformation, such as the Wavelet Transform (WT). Prior to feature transformation, various **Regions of Interest (ROI)**. Regions with low mean and variance in grayscale values of the various blocks are selected:

Figure 1: A scatter plot from Gillich and Lohweg, 2010 showing the decision boundaries for selection of viable regions of interest. The circles are classified as appropriate ROIs while the squares are discarded for non-homogeneity or high mean grayscale values.
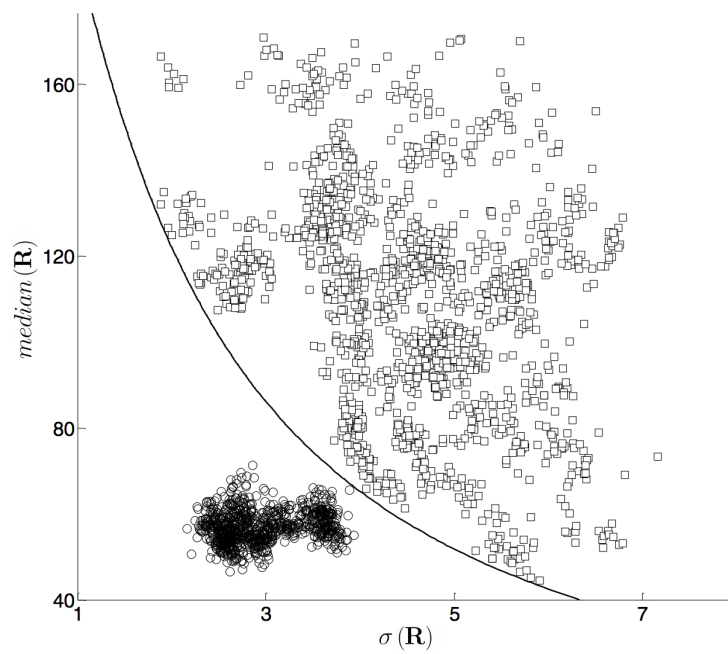
2 Related Work

3 Experiments

3.1 Validation of Model Performances

In their original paper, Gillich and Lohweg assert that 100% accuracy can be achieved as a result of judicious Region of Interest selection and the Shift Invariant Transform. To validate, we performed basic data processing, scaling the dataset so that each feature column k had $\mu_k = 0$ and a $\sigma_k = 1$:

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

df = pd.read_csv("data/dataset.csv", encoding="latin-1") # import the
    CSV dataset and save as a Pandas dataframe object

targets = df["class"].values # save the targets (Y) as a NumPy array of
    integers
features_df = df.loc[:, df.columns != "class"] # save everything else as
    the features (X)

features_names = features_df.columns.values # save the column names
features_matrix = features_df.values # convert X from a dataframe to a
    matrix (for our machine learning models)
scaled_features_matrix = StandardScaler().fit_transform(features_matrix)
    # scale the data so mean = 0, unit variance - important for many
    models
```

The cleaned `scaled_features_matrix` Numpy object can then be fed into a variety of different models:

```python
logistic_regression = LogisticRegression()
logistic_regression.fit(X_train, y_train)
training_predictions = logistic_regression.predict(X_train)
test_predictions = logistic_regression.predict(X_test)

# check accuracy of Logistic Regression
number_correct = np.sum(training_predictions == y_train)
number_correct_test = np.sum(test_predictions == y_test)
```

3.2 Sample Size Stress Testing

3.3 Implementation of Frog Boiling Attack (FBA)

The Frog Boiling Attack is most effective during business contexts where online learning occurs (when data arrives as a stream and becomes ingested sequentially, with the fitted model updated after each new data point). The attack relies upon the concept of template drift[**?**], where the model's fitted "template" of an authentic banknote data sample is corrupted over time by a stream of compromised data points. The overall implementation, implemented over N steps, relies upon the update equation

$$F_{new} = F_{old} + (i-1)\delta(x) \tag{1}$$

where

$$\delta(x) = \frac{\bar{x}_{target} - \bar{x}_{original}}{N} \tag{2}$$

3.3.1 FBA Countermeasures and Considerations

When discussing potential countermeasures to FBA, a tradeoff between responsiveness and security must be established. An obvious, yet sometimes unrealistic, method of prevention is to fully disable online learning completely, and "freeze" the authentication classifier. In many businesses cases, disabling online learning may result in a significant decline in model performance, especially if the feature space is exceptionally dynamic (ie. consumer entertainment and purchasing behavior, which is often a function of seasonality and inherent "drift").

4 Future Work

References

[1] Eugen Gillich and Volker Lohweg. *Banknote Authentication*. inIT - Institute Industrial IT, Department of Electrical Engineering and Computer Science, Ostwestfalen-Lippe University of Applied Sciences November 2010

[2] Van Renesse, R.L.: Optical Document Security, 3rd edn., Artehouse Boston/London (2005)

[3] Z. Wang, A. Serwadda, K.Balagani, and Vir V. Phoha, "Transforming animals in a Cyber-Behavioral Biometric Menagerie with Frog-Boiling Attacks" in *The IEEE Fifth International Conference on Biometrics: Theory, Application and Systems (BTAS 2012)*, Washington DC, 2012