Counterfeit Banknote Classification

Katherine Kuenster

Department of Computer Science and Mathematics, Santa Clara University
500 El Camino Real, Santa Clara, CA 95053, USA

kkuenster@scu.edu

***Abstract*— The increase of fake currency flowing through the global economy has increased significantly within the last few years. With the increased use of digital banking, there is a demand to be able to recognize authenticity immediately from scanning a banknote. By using the relationships between numerical image features, and their true classification, an algorithm can be developed to classify each input authentic or fake immediately after scanning, preventing the counterfeit bill from moving further through the economy. Many supervised machine learning algorithms can be used to achieve this goal. Support-Vector Machines (SVM), and logistic regression will be compared to find the best accuracy for the model.**

I. Introduction

In this project, I investigated how various scanned features of currency translate to either authentic or counterfeit. In order to achieve this, I took advantage of SVM and logistic regression in order to classify events into binary categories. The following is a documentation and comparison of the various supervised machine learning algorithms and their accuracy, precision, and recall scores.

II. Data

This data[1] was acquired through studies done working to identify ways that banknotes could be recognized as authentic or not through numerical values gathered via Wavelet Transformation (WT). The initial set consists of 1,372 events, where each event represents a singular scanned banknote. Using an industrial camera, 400x400 pixel images were captured for both real and counterfeit banknotes. Using WT imaging, numerical variables that are associated to the class of the note. The following features were collected via WTI to train the model to recognize the scan of a legitimate (0), versus illegitimate (1) banknote:

1. *Variance*
2. *Skew*
3. *Kurtosis*
4. *Entropy*

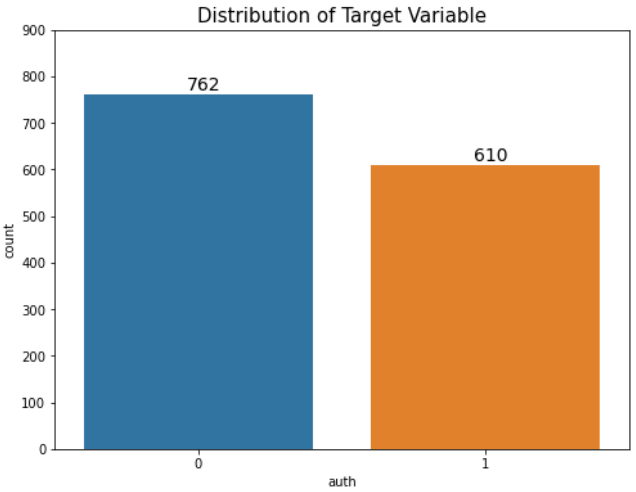

Fig. 1 A plot of the original dataset with the value count for each label, where 0 is the count of authentic bills in the dataset, and 1 is the count of counterfeit bills in the set

III. Normalization

The dataset has no missing values, however, as Fig. 1 shows, the dataset has an uneven distribution within the classes of the target variable; there are 152 more events for authentic banknotes (class 0), than counterfeit ones. This could lead to an overfitting for one class, and may lead to a model that is unable to recognize counterfeit notes as accurately as it recognizes authentic notes. To combat this, it was crucial to randomly

undersample events for class[0]. Normalization of data is achieved by dividing the difference of point X and the minimum value by the difference of the minimum and maximum values.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

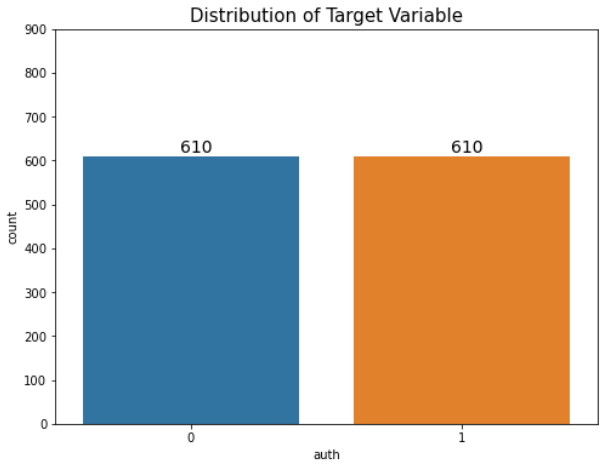

Fig. 2 A plot of the normalized dataset with the value count for each label, where 0 is the count of authentic bills in the dataset, and 1 is the count of counterfeit bills in the set

Fig. 2 shows the dataset to be used for training and testing our model. After randomly removing 152 events from class[0], the dataset is finalized with 1220 total events, 610 events per target variable class.

IV. Testing and Training Groups

In order to generate a model with classification capabilities, separating the dataset into a training and testing set such that the model performs on the testing set, based on the information processed from the training set. Data should be cleaned before splitting the dataset so that all events contain relevant data for the model to learn.

A typical split for datasets is to have a random 80% of the data assigned to the training set, and test the model on the remaining 20% in order to observe accuracy. With this split, there are 1,097 events training the model, which will be tested on the remaining 275 events.

V. Logistic Regression

Logistic regression is a supervised machine learning algorithm that is intended for datasets containing numerical input types, and a categorical target variable that contains two classes such that the relationship between the classes is binary. Logistic regression creates a linear boundary between classes in order to place future values into their correct classification.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

The hypothesis function, h(x), returns the predicted class that data point X belongs to. By comparing the predicted class to the actual class, we can find the accuracy, precision, and recall:

	Pred Negative	Pred Positive
True Negative	145	3
True Positive	0	127

By producing a confusion matrix, it is visible that the model has no false negatives, which means the model did not mistake any counterfeit notes for an authentic note, which is good for the purpose of this model, giving us a recall score of 1.0. With 3 false positives, the precision is 0.977. Overall, the accuracy is 0.989.

VI. Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning algorithm that is popular for classification and regression purposes. Specifically, SVM finds the optimal linear boundary of separation between two classes of data. Support Vectors are points in our space within each class that affect the positioning of the decision boundary.

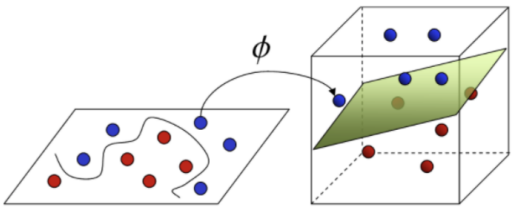

Fig. 3 A visual of SVM mapping from input space to feature space

Again, using the produced confusion matrix for SVM, the model accuracy, precision, and recall scores can be determined:

	Pred Negative	Pred Positive
True Negative	144	4
True Positive	1	126

SVM performed similarly to logistic regression with high scores for each measurement. With 1 false negative, and 3 false positives, SVM is nearly as accurate as logistic regression, but not quite. Accuracy comes to be 0.982, with a precision score of 0.969 and a recall score of 0.992. This means that roughly 96% of true positives were correctly classified. Furthermore, this informs that 98% of the events classified as positive were truly positive, (counterfeit).

VII. Conclusions

TABLE 1
Accuracy Report for Classification Algorithms

Method	Accuracy	Precision	Recall	F1
L.R.	0.989	0.977	1.0	0.988
SVM	0.982	0.969	0.992	0.981

Although SVM and logistic regression (L.R.) performed nearly the same, logistic regression performed slightly better, in all categories. We may decide we want to try and minimize our loss, however that is not necessary in this case as our accuracy is quite high, along with the model's ability to correctly classify authentic bills as authentic.

Acknowledgment

I would like to acknowledge the work of A. Bhatia, V. Kedia, A. Shroff, M. Kumar, B. K. Shah and their research that provided this dataset. I owe another thank you to Professor Ghosh at Santa Clara university for providing me the opportunity to further explore machine learning.

References

[1] A. Bhatia, V. Kedia, A. Shroff, M. Kumar, B. K. Shah and Aryan, "Fake Currency Detection with Machine Learning Algorithm and Image Processing," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*,2021,pp.755-760