Fake Currency Detection using Machine Learning Course:Machine Learning

1st Aleksandra Zdravkova 105635
DETI
University of Aveiro
Aveiro, Portugal
zdravkova@ua.pt
Course Instructor:Petia Georgieva

2nd Ana Atanasova 105648
DETI
University of Aveiro
Aveiro, Portugal
ana.atanasova@ua.pt
Course Instructor:Petia Georgieva

***Abstract*—The currency which is imitated with illegal sanction of state and government is counterfeit currency. Every country incorporates a number of security features for its currency security. Currency counterfeiting is always been a challenging term for financial system of any country. The problem of counterfeiting majorly affects the economical as well as financial growth of a country. In view of the problem various studies about counterfeit detection has been conducted using various techniques and variety of tools.**

Keywords-fake currency, machine learning, logistic regression, k-nearest neighbours,confusion matrix

I. Introduction

Almost every major company is under the threat of counterfeit money.Counterfeit money is currency produced without the legal sanction of the State or government, usually in a deliberate attempt to imitate that currency and so as to deceive its recipient.Producing or using counterfeit money is a form of fraud or forgery, and is illegal. The business of counterfeiting money is almost as old as money itself. Some of the bad and side effects that fake currency has on society is the reduction in the value of real money, and increase in prices, that leads to inflation, due to more money getting circulated in the economy. With the today's computers and technology, it became very easy and possible to make high quality counterfeit money that are hard to be recognized from real banknotes.

Machine learning techniques can be used to build tools which can help in this problem. We can make our computers to make and memorize patterns to help recognize between fake and real currency. Using these patterns and information, they will be able to classify a new currency bill as either fake or genuine.

II. Project Overview

Given the huge threat created by the usage of high quality counterfeit currency bills, it is desirable to investigate ways to try to distinguish them. Therefore we intend to develop a machine learning model that will do that. The model should be able to identify between fake and real currency with good accuracy. The training model will be made by using dataset with samples from both types of currencies. After the training and validation is completed,the model should be able to distinguish whether a new currency bill is real or fake.

III. Datasets

For this problem, we are using dataset for banknote authentication. The data is extracted from images that were taken for evaluation of an authentication procedure for money. Types of tools were used to extract the features from the images. The extracted features have these attribute information:

- Variance of the image transformed into wavelets
- The asymmetry of the image transformed into wavelets
- Kurtosis of the image transformed into wavelets
- Image entropy
- Class of the currency

The first four features are continuous and they define the characteristics of the currency, while the last feature is integer and has values 1/0. 0 is for the real banknotes and 1 is for the fake one.

This dataset contains total of 1372 samples, from which 762 samples belongs to fake notes and 610 samples belong real notes.

A. Distribution of Features in Banknotes in Dataset

Sometimes, a dataset can contain some features whose values tend to be near one number, but also to have a non-trivial number of smaller or larger values than that number. On Figure 1 and Figure 2 ,we plot a histogram for each feature of the dataset.

- From the histogram we can see that there are no features for which a sample has a very large of very small number of value than the rest of the samples.
- These two features,variance and skewness, are pretty much evenly distributed.
- Kurtosis has positively skewed distribution, and most of the samples have value less than 3. For those ones, the distribution produces less outsiders(outliers) than the normal distribution with kurtosis value 3.
- Entropy has negatively skwed distribution, which means more samples have higher entropy and that implies to that more images have higher contrast.

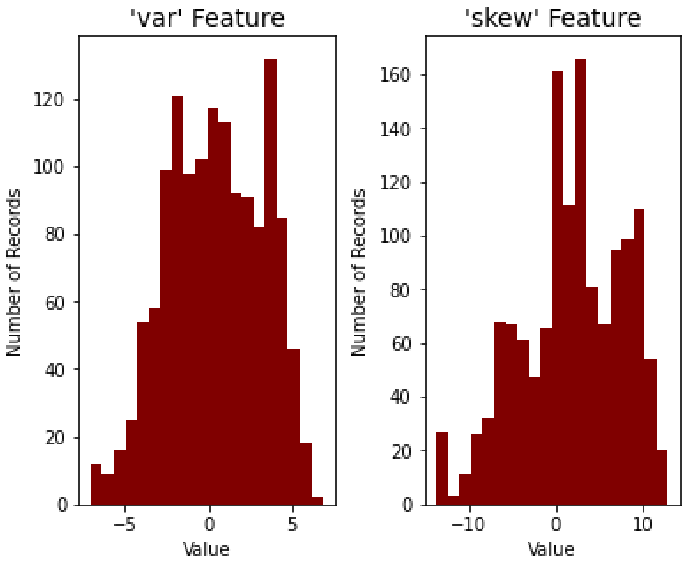


Fig. 1. Histogram for Variance and Skewness Features

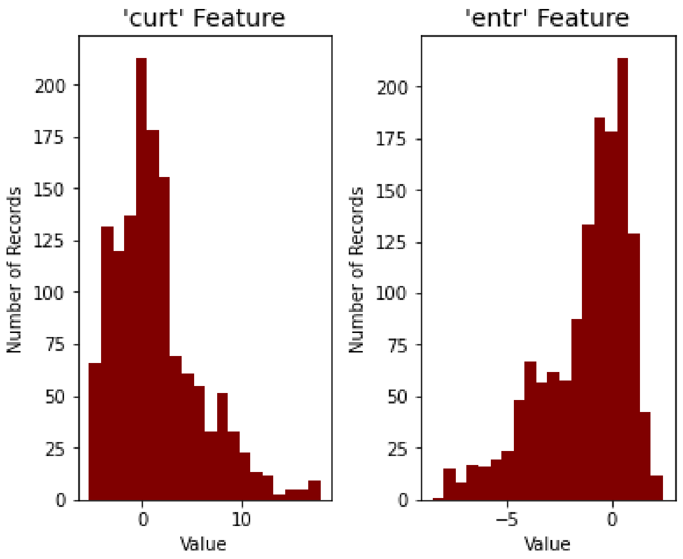


Fig. 2. Histogram for Kurtosis and Entropy Features

B. Scatter Analysis

For this analysis we created scatter plots with the four features, and we use different color for samples who belong to class 1 and class 0. The different color in the scatter would help us easily identify the real currency(class 0) from the fake one(class 1).

From the plots from the scatter analasys we can see:

- Skewness-Variance pair most strongly separates the two classes of currencies and draws a boundary between them.
- Entropy-Kurtosis pair least strongly separates the two classes of currencies and with this pair algorithms will have hard time in making difference between real and fake money, if we only use these two features.

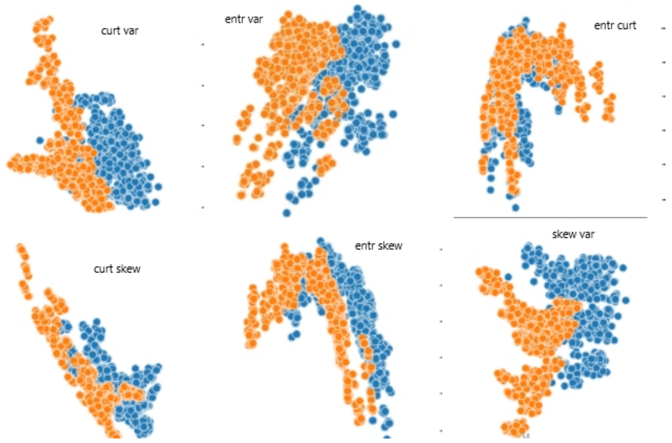


Fig. 3. Scatter plot analysis

- Variance is notable feature that draws line between fake and real money. Real currency has low variance, while the fake ones high variance.
- When it comes to the entropy, it is not that important feature in the class-separability.

IV. Data Processing

After finishing the Scatter Analysis, we need to check if our data is balanced or not. If the data is perfectly balanced, we can go to the next step and do your binary classification, but if the data is still not perfect we need to balance it again and try to make it perfect.

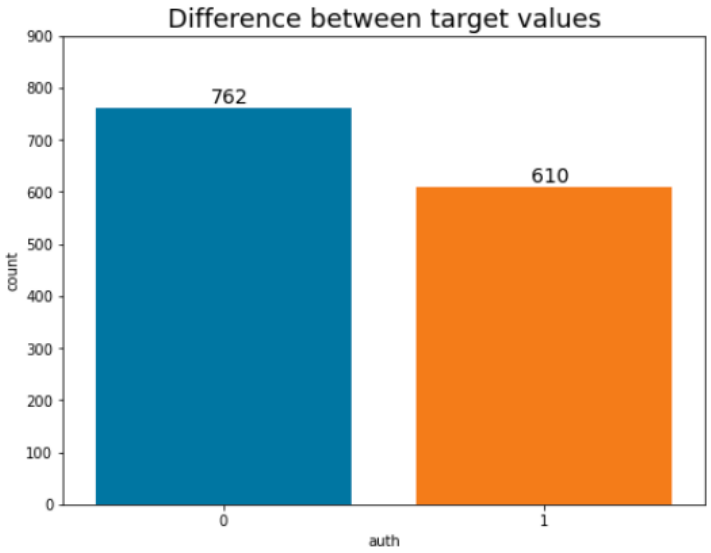


Fig. 4. Balancing data set

As we can see from the graph above, our data set is quite good balanced, but it's still not perfect. So because of that, we can do two things.

- Random under-sampling- to drop randomly a number of instances of the target that have more, in this case from real banknotes. It involves randomly selecting examples from the majority class and deleting them from the training dataset.

- Random oversampling- to create more new data for the target that is under-represented (fake banknotes). It involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset.

In our case, we chose to do random under-sampling and to lower the data of real banknotes by 152 observations. So after that the data is perfectly balanced with 610 observations for real an 610 observations for fake banknotes.

V. Test and Training Sets

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, and when it is used for classification and the dataset is not balanced. That is why in previous step we balanced our data to perfection.

Usually, when we use train and set data, We usually split the data around 20 percents-80 percents between testing and training stages.

The point of doing this, and using test and training sets, is to measure the accuracy of your model, which is very important.

- You train the model using the training set. Train the model means create the model.
- You test the model using the testing set. Test the model means test the accuracy of the model.

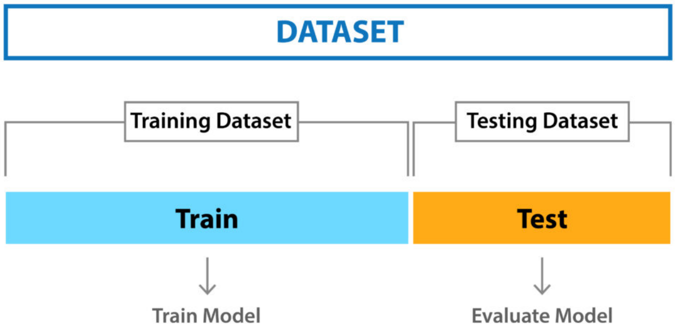


Fig. 5. Train and Test Set

In our project, we divided the data and we got 976 for training set, and 244 for test set.

VI. Algorithms and Techniques

For this project one of the algorithms that we used is Logistic Regression

A. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model.In statistics is used to model the probability of a certain class or event existing such as pass/fail, win/lose etc. In our case, we used the Logistic Regression for training and testing the model with first fitting the data into the model.

B. Types of Logistic Regression

- Binary Logistic Regression- is the statistical technique used to predict the relationship between predictors (our independent variables) and a predicted variable (the dependent variable) where the dependent variable is binary. The categorical response has only two 2 possible outcomes. In our case that will be real banknote and fake banknote.
- Multinomial Logistic Regression- is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership.
- Ordinal Logistic Regression- assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories.Three or more categories with ordering.

Last, but not least, we used the Logistic Regression to train the accuracy of our model and it achieved accuracy of 98.36

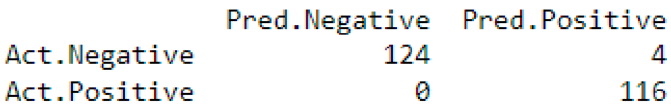
C. Confusion Matrix for Logistic Regression

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

The basic terms, which are whole numbers in confusion matrix are:

- TP-true positive- These are cases in which we predicted as fake banknotes , and they are really fake. In our case that is 116
- TN-true negative- We predicted that the banknotes are real, and they are actually real. In our case that is 124.
- FP-false positive- We predicted that the banknotes are fake, but in fact they are real.Also known as a ”Type I error. In our case that is 4.

- FN-false negative- We predict that the banknotes are real, but they actually are fake. Also known as a ”Type II error. In our case that is 0.



```
              Pred.Negative  Pred.Positive
Act.Negative            124              4
Act.Positive              0            116
```

Fig. 6. Confusion Matrix

D. K-Nearest Neighbours Classifier

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. The output depends on whether k-NN is used for classification or regression.Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

K-Nearest Neighbours,in a way, is a class of lazy learners which classifies the data by looking an its neighbours and assigning weighs to them, where the one who are closest to the given point have much more saying than the others points. Some of the parameters that KNN uses are:

- N-neighbours:The number of neighbours that are used in the classification
- Weights: The degree of the influence each data point has it
- Algorithm: Some kind of algorithm that is used to compute the nearest neighbours to the given point.

Some of the advantages KNN has are:

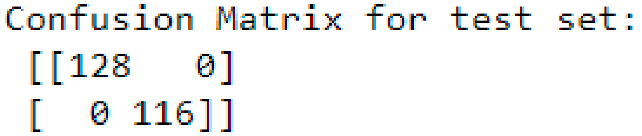
- No training period. Because it’s called lazy learner,KNN does not learn anything in the training period. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training.
- It is very easy to implement
- New data can be added seamlessly, which won’t impact on the accuracy of the algorithm.

Disadvantages of KNN are:

- Doesn’t work well with large datasets, because distance between points is huge and that degrades the preformance of the algorithm.
- Doesn’t work well with high dimensions,because it becomes difficult for the algorithm to calculate the distance in each dimension.
- Sensitive to noisy data, missing values and outliers.

E. Confusion Matrix for KNN Classifier

- TP-true positive- These are cases in which we predicted as fake banknotes , and they are really fake. In our case that is 116
- TN-true negative- We predicted that the banknotes are real, and they are actually real. In our case that is 128.
- FP-false positive- We predicted that the banknotes are fake, but in fact they are real.Also known as a ”Type I error. In our case that is 0.
- FN-false negative- We predict that the banknotes are real, but they actually are fake. Also known as a ”Type II error. In our case that is 0.



```
Confusion Matrix for test set:
 [[128   0]
 [  0 116]]
```

Fig. 7. Confusion Matrix for KNN Classifier

VII. Improvement

Even though we had good accuracy for detecting the fake currency with this model,there is always chance of improvement and making it better, by having lager datasets which can capture more patterns and features and exploring other algorithms and techniques which we did not use in this project. Also another solution that can be used to solve this model is the use of deep learning. A convolution neural network can be trained,directly over the captured images, to build a classification model.

VIII. Conclusion

In this project, we accomplished our goal by successfully implementing Logistic Regression and KNN Classification. Using already existing data for real and fake banknotes, we were able to tell and make a difference between them. Improving and perfecting this model can be really useful, because nowadays fake banknotes are more and more used.

References

https://github.com/dudeanurag/Fake-Currency-Identification

Counterfeit currency detection techniques: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8443015

Confusion Matrix: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Logistic Regression: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc Test and Training Sets https://www.w3schools.com/python/python$_{m}l_{t}rain_{t}est.asp$