Statistical Machine Learning for Counterfeiting Money Detection.

Jean de Dieu Murera Gisa

June 10, 2019

Abstract

The counterfeit money is termed as the imitation of the nation currency generated without a legal sanction of the state or autonomous government. This illegal act of money forgery is already getting global threat and significantly increasing as technology, computerization, and high definition photocopy and printer technology do and make it quite possible in return for individuals without sophisticated training and intelligence to easily currency fraudulent. Many counterfeited currency ill-effects are massively prevailed in country economy including the money depreciation and devaluation, inflation, shrinkage of currency confidence on global market and heavy hurting of the domestic business performance. Therefore the modern strong anti-counterfeiting measures and detection techniques are put in place to overcome these effects. And as time goes on, Machine Learning and Artificial Intelligence becomes one of the invaluable approach among those anti-counterfeiting techniques due to its both straightforwardness in computational implementation and ability to output the trusted optimal results. The crucial intent of this paper is to quite employ the several thirteen statistical machine learning algorithms on the banknotes authentication data to classify, detect and predict the counterfeited and genuine banknotes circulating in the country economy. Besides, the Mathews's Correlation Coefficient has been used as the model performance metric to measure the goodness of predictive performance of each learning machines employed. The adaboost algorithm has been selected and presented as the superior optimal model for the task with maximum correlation (agreement) between the actual and predicted banknotes classes.

Keywords: Banknote Authentication dataset, Statistical Machine Learning Algorithms, Patterns Recognition and Classification, Mathews' Correlation Coefficient.

1 Introduction

Normally, a central bank, reserve bank or monetary authority of an autonomous state is one and only institution which has the right sole authority to issue, deliver, change and design the banknotes circulating in the economy. However, the counterfeiting banknotes is sufficiently prevalent worldwide that has been called the world second oldest profession (Witschorik, 2000). The business of money fraudulent is as old as money itself, the coinage of money began in the Greek city of Lydia around 600B.C. Before the introduction of paper money, the most prevalent technique of currency forging involved combining base metals with pure gold and/or silver which is quite difficult today because of scarcity and the cost of such precious metals. But in this era of paper banknotes, the millions of fake notes are passed over the retails counters and the majority are not identified. And nowadays, The different central banks of different countries have adopted the numerous anti-counterfeiting techniques to decrease as well as to stop those forged notes called Superdollars with high quality and likeness to the authentic dollars. Among them including the sophisticated banknotes design and fine detail with raised intaglio printing on bills as shown on Figure (1). This would allow non-experts to easily spot banknotes forgeries.

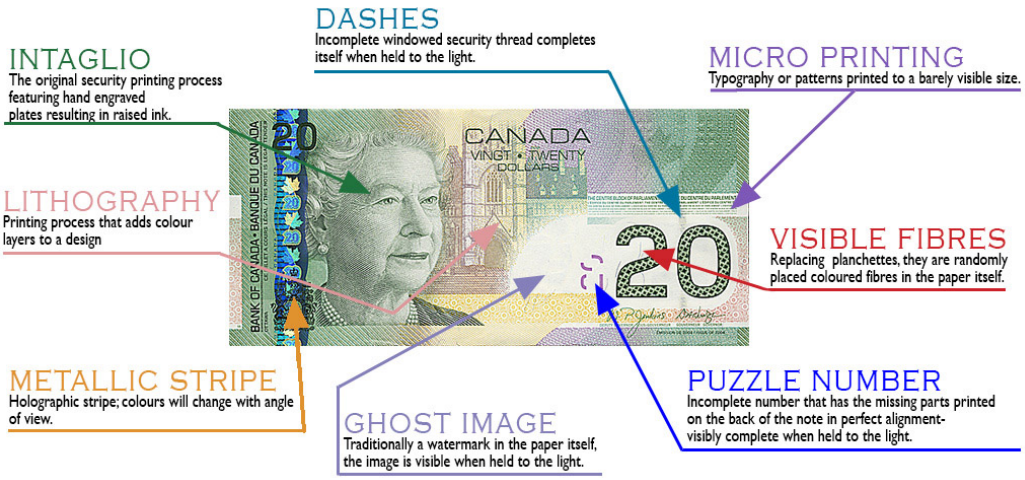


Figure 1: Sophisticated bill design and raising of intaglio printing. (wikipedia.org, 2015)

However, the talented counterfeiters are likely to forge and multiply the banknotes look quite similar to genuine ones to circulate in the world economy with the unknown and disputed sources (Murthy *et al.*, 2016). The Figures (2a) and (4b) present the counterfeited currencies and their similarities to the authentic ones.

(a) Forged Banknotes (wikipedia.org, 2015)

(b) Forged Banknotes

Today's advancement of technology and printing industries has accelerating the rapid growth in the counterfeiting money. Due to this fact, the myriad of anti-fraudulent methods have been put in place to decline the counterfeited banknotes and their associated ill-effects on the community and on the country economy at large. The various techniques have been adopted to detect the forged currencies. Among them the artificial intelligence and machine learning technique can't be left behind in the task as long as it solve significantly and surprisingly the current complex problems beyond the human expectations.

2 Machine Learning and Performance Metrics

This section clarifies all steps required to build the proper machine learning project and effective usage of Learning algorithms especially supervised ML to tackle the considered problem in different tasks assigned. It also presents one of the evaluation metrics for the supervised classification learning machines used in this piece of work.

2.0.1 Statistical Learning Machine

Machine Learning (ML) is the science of generating the computations and algorithms that allow software applications to become more accurate in predicting outcomes without being explicitly programmed (Pant, 2015) and simplify the work

of humans. ML is powered by data and enables computers to learn from them to make tremendous automatic data-based decision and prediction. Having said that ML algorithm needs data to deliver insightful information, it is not only the data trained but also the size of data considered. With the exponential increase in data, ML become more valuable and powerful tool in world problem solving platform.

Machine Learning is often categorized into supervised, unsupervised and reinforcement learning, but the supervised one has been used to distinguish valid and counterfeit banknote. Here below is the workflow of supervised learning machine algorithms used in this piece of work.

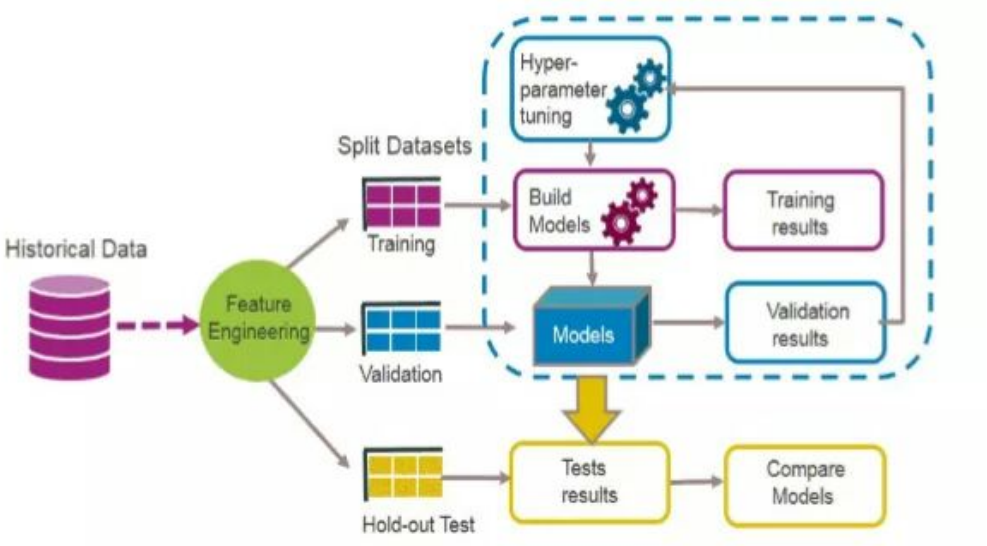


Figure 3: WorkFlow of ML algorthms used.
(Pant, 2015)

Simply, the Figure (3) above summarizes the taken processes of data gathering, preparation and pre-processing, model building, testing and validation in the almost relevant supervised machine learning project.

2.1 Predictive Models Performance Evaluation Metrics

Deciding on the right evaluation metric for the machine learning algorithm for either classification and regression models is quite crucial and relevant to the success of any machine learning project. So, for the excellency and such success of any machine learning based project, it is relevant to be cautious in choosing the evaluation metrics to be used for clear model performance measurement basing on the related tasks. Then, for learning machine classification task, there are

many classifier (predictor) evaluation metrics which differ from one field to another according to the expected goals and targets to achieve. Basing on this, we have chosen to follow the excellent way in selecting one of the best model performance metric untitled "Mathews' correlation Coefficient" for better evaluation of the model performance that will lead to the undoubtedly selection of superior efficiency and efficacy algorithm relative to others being used.

2.1.1 Matthews Correlation Coefficient (MCC)

The different existing data Scientists claim that this model performance metric to be the most informative single score to trace and establish the quality of binary classifier than other confusion matrix measures (such F1 score(harmonic mean of precision and recall) and Accuracy) (Boughorbel *et al.*, 2017). Since it puts into consideration the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives) (Gorodkin, 2004), Mathews's correlation coefficient has been selected and chosen as the relevant metric to be used in this study in avoidance of the drawback and misleading of some of the other metrics. In addition, this metric is good for our dataset since it is unbalanced of negatives and positives data features in which the other metrics including accuracy and F1 Score couldn't probably estimate well if the predictor is accurate as expected in this case.

2.1.2 Mathematical Notation of MCC

Mathematically, Matthews Correlation Coefficient(MCC) of binary classifier θ is expressed as below:

$$\mathrm{MCC}(\theta) = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

Where, TP is True Positive(TP), TN, True Negative(TN), FP, False Positive(FP) and FN, False Negative(FN) all obtaining in confusion matrix which gives a better summary of the performance of classification algorithms. Note that, the metric is not defined if any of the sums in denominator of equation (1) is zero.

2.1.3 MCC Based Model Quality Interpretation

The MCC metric takes values in the interval [1, +1], with +1 indicating a complete strong perfect agreement, 1 a complete perfect disagreement between prediction and observed values, and 0 showing that the prediction was uncorrelated with the ground truth and draw the idea that the model prediction is no better than random prediction.

3 Methodology

This section begins by first describing data set used and presenting the different repositories in which it has been extracted and sourced. Then, explain the systematic approaches carried out to detect, classify, and predict the banknotes whether they genuine or not using different thirteen statistical learning machine algorithms. And later, the model performance metric.

3.1 Source and the Description of Data set

The Banknote authentication data set is real-valued continuous multivariate set of data which has been extracted from various genuine and forged banknote-like specimens images. For digitization, the high definition industrial camera and fundamental image processing tool (Wavelet Transform Tool) were used to produce such data set of 1372 observations and 5 attributes with the intention of evaluating the authenticity of banknotes. The data set is publicly available from various ML repositories including data hub [1], Kaggle Datasets [2] and University of California Irvine (UCI), center for Machine Learning and intelligent systems repository. [3] The below Table (1) is presenting the details on data set employed in this paper.

Table 1: Banknotes Authentication data set Description.

Data Set Characteristics	Multivariate	Observational Instances	1372
Attributes Characteristics	Continuous	Number of Attributes	5
Date Donated	16/04/2013	Missing Values	NA

3.2 Experimental Systematic Approach Undertaken

3.2.1 Statistical Tool, Libraries and Packages

The R programming language were used as statistical tool to implement the thirteen machine learning algorithms to detect, classify and predict the prevalence of banknotes classes. R language is an open-source software for statistical analysis, computing, graphics representations and reporting (Fox and Andersen, 2005). R is freely available under the GNU, general public and pre-compiled binary versions are provided for various operating systems like Unix, Windows and Mac.

[1] https://datahub.io/machine-learning/banknote-authentication
[2] https://www.kaggle.com/sarathp97/banknote/version/1
[3] http://archive.ics.uci.edu/ml/datasets/banknote+authentication

The statistical R libraries and packages relevant for this paper are differ from their assigned specific tasks and functions. Here below are those R libraries utilized to implement the thirteen statistical learning machines to detect fraudulent banknotes.

- **The package caret:**[4] It stands for Classification and Regression Training (Kuhn *et al.*, 2008). It accommodates and integrates more than two hundreds and nearly every major machine learning algorithms available in R language. This library has been widely used in this paper since it loads all thirteen algorithms used for the experiment.
- **The tidyverse package:**[5] The tidyverse is an opinionated collection of R packages designed for data manipulation, exploration and visualization. All hosted packages share an underlying design philosophy, grammar, and data structures (Wickham).

To implement each and every supervised statistical machine learning algorithm considered to detect, classify and predict the prevalence classes of banknotes, we have used the function *train()* supported by library *caret* and mentioned the recognized name of the algorithm in argument *method()*.

3.2.2 Experiment Set Up and Data Pre-Processing

The employed Banknote authentication data set has been well prepared, pre-processed and splitted into training and independent test data sets before heading to the models estimation. The predictive performance of each learning machine model has been estimated by using the algorithmic syntaxes supported by *caret* library. It has several functions that attempt to streamline the predictive model building and evaluation processes, as well as feature selection, positioning and scaling, data splitting, pre-processing, and variable importance estimation. The fundamental and principal tool in such library is the ***train()*** function which has been used to train each model performance on splitted train set of fixed sample dimension of $(1098, 5)$. And it has been used to select the optimal model across the hyper-parameter tuned parameters through utilization of 10-fold cross validation repeatedly three times as re-sampling method. Therefore, the models have been validated on test data set of sample dimension of $(274, 5)$ to predict the prevalence banknotes classes. The author can't leave behind the fact of setting pseu-random number generator for the avoidance of result variability in order to obtain the nearly trusted experimental results.

[4] http://CRAN.R-project.org/package=caret
[5] https://www.tidyverse.org/

More formally, the *train()* Pseudocode is used for simplicity and easy understanding of each model algorithm infrastructure and program skeleton to implement all considered statistical machine learning models which were compiled without errors.

Algorithm 1 :Model Training and Hyper-parameter Tuning

Define sets of model parameter to evaluate
for *each parameter set* **do**
  for *each re-sampling iteration* **do**
    Hold-out specific samples
    [Optional] Pre-process the data
    Fit the model on the remainder
    Predict the hold-out samples
  end for
  Calculate the average performance across hold-out predictions
end for
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set

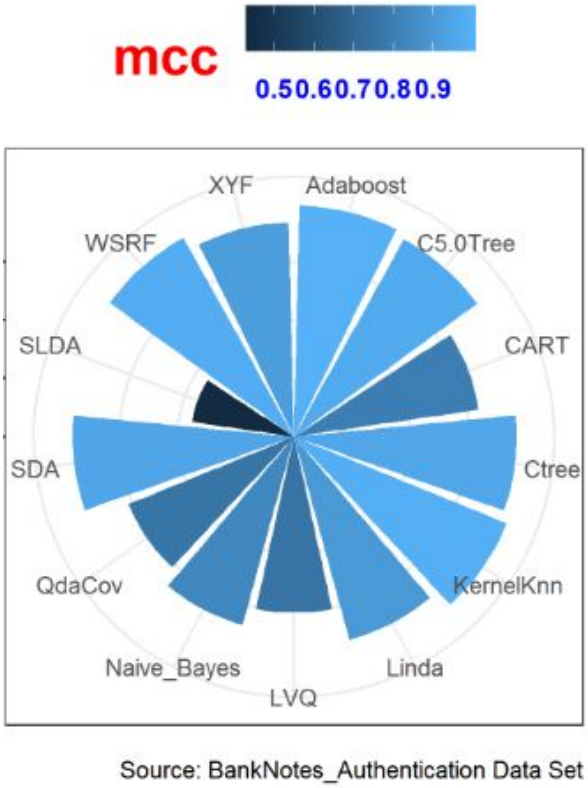
3.3 Optimal Model Selection and Performance Evaluation

The banknotes classification into genuine or forged notes was conducted using the thirteen supervised learning machine algorithms through the aforementioned systematic approaches. The predictive built models have been evaluated to quite measure the model's quality and efficacy performance by using the Mathew's correlation coefficient (mcc). It indicates the agreement between the actual(observed) and predicted prevalence classes in binary pattern recognition and classification tasks. The predictive model performance were recorded and presented in table format. Then a model with maximum correlation coefficient has been selected and chosen as an optimal model due to the fact that it represents the perfect prediction relative to others.

4 Result and Discussion

This section represents the experimental results obtained from training thirteen classification supervised algorithms on banknote authentication data set in order to significantly detect the counterfeited money circulating in the economy. The quality of the model has been presented in terms of its Mathew's correlation coefficient.

Firstly, the predictive model comparative results is presented, later the optimal model basing on its maximum agreement between observable and predicted banknotes prevalence classes. The Table (4b) and Figure (4a) are presenting the comparative model performance with Mathews correlation coefficients.



(a) Model Performance Pie Chart.

(b) Model Performance by mcc metric

SN	ML Model	mcc
1	SLDA	0.4417760
2	LVQ	0.7562242
3	QdaCov	0.7668727
4	CART	0.8003889
5	Naive_Bayes	0.8375062
6	Linda	0.9044117
7	XYF	0.9204091
8	SDA	0.9498330
9	Ctree	0.9556730
10	C5.0Tree	0.9709653
11	WSRF	0.9778811
12	KernelKnn	0.9853554
13	Adaboost	0.9926452

Figure 4: Classifiers Comparative performance on Banknotes Authentication data.

The comparison of experimental measurement presented from Figure (4) gives the right decision to select the Adaptive Boosting meta algorithm as the superior optimal model relative to others, all trained on 80% training and 20% testing of Banknotes authentication dataset. Since it is used in conjunction with other types of learning algorithms to improve performance, it won the detection, classification and prediction of counterfeited banknotes circulating in economy. In addition, it is often referred to as the best classifier due to its achievement of optimal performance of the 99.26% agreement between the actual and predicted banknotes prevailed classes. Besides, adaboost model predicted 151 forged banknotes which represents 55.1% and genuine notes of 123 represents 44.9% on the validation data set.

5 Conclusion

This study considered 13 different statistical machine learning algorithms which have been carefully selected from the set of supervised parametric and/or non-parametric models. To select the best optimal model, the *caret* R package has been widely used to train and tune our models. The adaboost model takes the lead in accuracy relative to others with the maximum agreement (Mathews' correlation coefficient) between the actual and predicted banknotes prevailed classes.

References

C. A. Witschorik, *System and method for the detection of counterfeit currency*, 2000, http://journal.frontiersin.org/article/10.3389/fnint.2017.00021/full, US Patent 6,131,718.

wikipedia.org, *Towards Data Science*, 2015, 2–4.

S. Murthy, J. Kurumathur and B. R. Reddy, 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1–6.

A. Pant, *Towards Data Science*, 2015, 2–4.

S. Boughorbel, F. Jarray and M. El-Anbari, *PloS one*, 2017, **12**, e0177678.

J. Gorodkin, *Computational biology and chemistry*, 2004, **28**, 367–374.

J. Fox and R. Andersen, *Department of Sociology, McMaster University*, 2005, 2–4.

M. Kuhn *et al.*, *Journal of statistical software*, 2008, **28**, 1–26.

H. Wickham, *URL https://CRAN. R-project. org/package= tidyverse. R package version*, **1**, 51.