Deception Detection for MU3D using Machine Learning Algorithms

Kun Bu and Kandethody Ramachandran

University of South Florida kunbu@usf.edu, ram@usf.edu

Abstract

The purpose of this work is to detect people lying using different machine learning algorithms via analyzing the micro-expressions when people speak, and to conclude which algorithm gives the best prediction. The dataset that we used to fit into the machine learning models is Miami University Deception Detection Database (MU3D). MU3D is a free resource containing 320 videos of Black and White targets, female and male, telling truths and lies. We fit the MU3D video level dataset into difference machine learning algorithms, and give conclusion for the accuracy of prediction.

Keywords: machine learning classification, deception detection, exploratory data analysis (EDA), support vector machine, random forest.

1. Introduction

Traditional lie detection machine is a polygraph, which can provide people with an averaging accuracy between 58% to 90%. With 90% accuracy, it seems to do a very good job on detecting lying, however, with 58% accuracy, we can hardly have much confidence to say this person is lying. In other words, the polygraph test is easy to pass for those well-trained people (ie. company spies or country spies). Even ordinary people who search for the word "polygraph" online, the next searching suggestion would be "How to Pass a Polygraph Test?" Since the polygraph operating principle is to detect lies by looking for signs of an examinee's physiological changes. Once the examinee lies, it puts a blip on the polygraph machine that serves as a signature of that examinee's lies. Besides, polygraph test is a time-based test that only captures the examinee's body reaction in each specific question, which means the examinees themselves know that they're being tested whether they are lying. Therefore, polygraphs are not useful for those underground and secret cases. Therefore, artificial intelligence (AI) approaches come to scientist's minds. Why don't we just detect lying by applying machine learning algorithms to see if the accuracy of deception detection would be improved.

The Miami University Deception Detection Database (MU3D) is a free resource containing 320 videos of Black and White targets, female and male, telling truths and lies. Eighty (20 Black female, 20 Black male, 20 White female, and 20 White male) targets were recorded speaking honestly and dishonestly about their social relationships. Each target generated four different videos (i.e., positive truth, negative truth, positive lie, negative lie), yielding 320 videos fully crossing target race, target gender, statement valence, and statement veracity. The stimuli and an information codebook can be accessed free of charge for academic research purposes from http://hdl.handle.net/2374.MIA/6067. In the previous studies of MU3D, scholars conducted research using standardized stimuli that can aid in building comprehensive theories of interpersonal sensitivity, enhance replication among labs, facilitate the use of signal detection analyses, and promote consideration of race, gender, and their interactive effects in deception detection research.

This paper is organized as follows. In section 2, we will talk about the related work in deception detection; section 3 provides the exploratory data analysis (EDA) of dataset to solving deception detection. Section 4 explains our experimental setup. In section 5 and 6, we discuss our results and drawbacks respectively. And finally, conclude with future work in section 7.

2. Related Work

As a widely studied phenomenon in many disciplines, deception, in psychology, is defined as an act that is intended to foster in another person a belief or understanding which the deceiver considers false [1]. Previous work on deception detection has focused on a combination of different factors including verbal and non-verbal aspects. Text / audio only approaches alone using RNN or LSTM architecture were able to achieve only moderate amount of accuracy 76% - 84% [2] . Micro-expression only approaches achieved higher accuracy of 77% - 88% [2]. Abouelenien et al., Akoglu et al., and Ott et al. have been extensively studied the challenging task for human deception detection. In their previous work, the accuracy of deception detection that are predicted by machines are approximately as high as 90%, where is a great achievement compared with the accuracy that predicted by human [3]. The task introduced by Ott et al. provides an ideal sandbox to understand human predictions with assistance from machine learning models [4].

Lai et al. summarize related work in two areas to put the deception detection work in context: interpretable machine learning and deception and misinformation [4]. They claim that Machine learning models remain as black boxes despite wide adoption. Blindly following machine predictions may lead to dire repercussions, especially in scenarios such as medical diagnosis and justice systems [5]. Therefore, improving their accuracy and interpretability has become the new trend in ML studies [6]. To improve the ML preformance, Chen et al. investigates the performance of combining support vector machines (SVM) and various feature selection strategies, for example, they applied both filter-type approaches and wrapper-type methods for feature selection [7]. Xiong et al. introduce a new large margin classifier, named SVM/LDA, which is an extension of support vector machine by incorporating some global information about the data, and their new formulation is expected to perform better or similar to standard SVM classifiers, and this assertion is empirically verified using several artificial and real-life benchmark data sets [8].

3. Exploratory Data Analysis

As mentioned in the introduction part, the MU3D is collected by recording 80 targets speaking honestly and dishonestly about

Video-Level Variables	Description
VideoID	ID associated with the video.
Valence	Indicates whether the statement in the video is negative or positive: value of 0 a indicates negative statement, value of 1 indicates a positive statement.
Veracity	Indicates whether the statement in the video is a truth or a lie: value of 0 indicates a lie, value of 1 indicates a truth.
Sex	Indicates target's sex: value of 0 indicates a female target, value of 1 indicates a male target.
Race	Indicates target's race: value of 0 indicates a Black target, value of 1 indicates a White target.
VidLength_ms	Indicates length of the video in milliseconds.
VidLength_sec	Indicates length of the video in seconds.
WordCount	Indicates the number of words contained in the full transcription of the video.
Accuracy	Indicates average accuracy (i.e., proportion correct) across raters who viewed the video.
TruthProp	Indicates average truth proportion (i.e., proportion of truth responses) across raters who viewed the video.
Attractive	Indicates average attractiveness ratings (measured on a scale ranging from 1 "Not at all" to 7 "Extremely") across raters who viewed the video.
Trustworthy	Indicates average trustworthiness ratings (measured on a scale ranging from 1 "Not at all" to 7 "Extremely") across raters who viewed the video.
Anxious	Indicates average anxiousness ratings (measured on a scale ranging from 1 "Not at all" to 7 "Extremely") across raters who viewed the video.
Transcription	Full transcription of the video.

Table 1: MU3D Video level database variables description [9]

their social relationships. The dataset was divide into two parts: video level and target level. In the video level dataset, information such as the valence, indicating whether the statement in the video is negative or positive; VidLength_ms and VidLength_sec indicates the length of video in millisecond and seconds, respectively. There are a total number of fourteen variables in the video level dataset, a short variables description is shown in the Table 1.

Excepted VideoID and the last variable Transcription, a boxplot was created for all other variables by Veracity. As shown in Figure 1, only the variable Accuracy has difference in Veracity, which is hard for us to start choosing features to train a classification prediction model. A correlation scatter plot was shown in Figure 2, it indicates VidLength_ms and VidLength_sec, variable TruthProp and Trustworthy, variable Accuracy and TruthProp are highly correlated, however, the effect here is similar to that of multicollinearity in linear regression. Our learned model may not be particularly stable against small variations in the training set, because different weight vectors will have similar outputs. The training set predictions, though, will be fairly stable, and so will test predictions if they come from the same distribution. We will applied a Principle Component Analysis (PCA) to help reduce the dimensions.

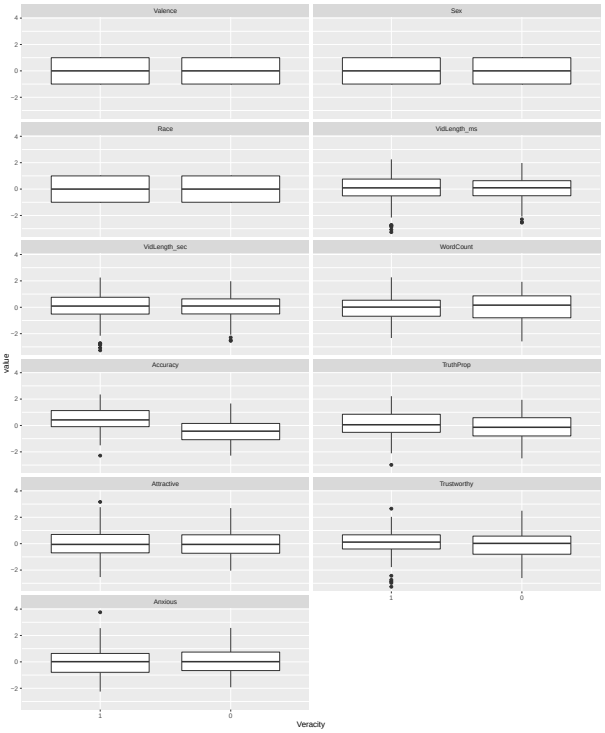

Figure 1: Boxplot for video level dataset variables

4. Method

The next step is to fit MU3D into a machine learning model and see how the computer performs on detecting lies. We will train different models based on the data properties. Three machine learning models including Support vector machine (SVM) and Binary Logistic Regression (BLR) and Random forest (RF) are applied to predict the deception.

Support Vector Machines are based on a decision plane concept that defines decision boundaries. A decision plane is a separation plane between a set of objects with different types of membership. SVM is a supervised learning method used to perform binary classification on data. In our case, we have exactly two classes : Lies or Truth. Besides, SVM can deal with real valued features, which means there are no categorical variables in the data, such as our dataset above, all of the features are from the facial landmarks coordinates, they are all numerical numbers, which are much fittable by using SVM. What's more, the SVM can perform well on a large number of features, for example, it works with ten, hundreds and thousands of features. In our dataset, we have a large number of features which motivates me to choose SVM. Another reason I would like to mention here is that SVM has simple decision boundaries, indicating that there are no issues with over fitting. The SVM can be defined as linear classifiers under the following two assumptions:

- The distance from the SVM's classification boundary to the nearest data point should be as large as possible;
- The support vectors are the most useful data points because they are the ones most likely to be incorrectly classified.

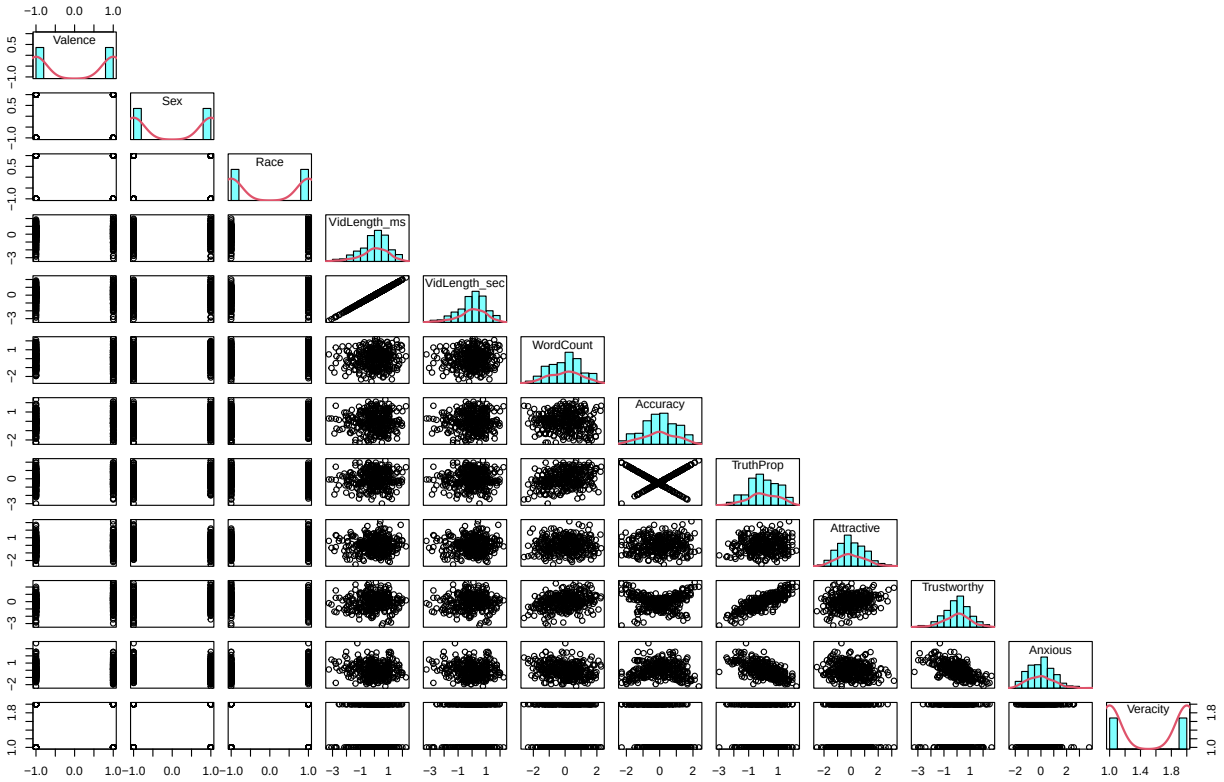

Figure 2: Correlation scatter plot for video level dataset variables

Thus, SVMs can be very efficient in predicting a data point's class. In a dimension less than two, the boundary is linearly separable classes in the plane; and in more than two dimensions, the boundary is known as a hyper-plane.

BLR is a statistical tool that classifies local node behavior to either malicious or benign. BLR has two stages: training and evaluation. At the training stage it uses node behavior from both benign and malicious node activity and derives a detection module. At the evaluation phase, data that was not used in the training stage, is used to evaluate the detection model. Regression analysis is a process that estimates the probability of the target variable given some linear combination of the predictors. Binary logistic regression (LR) is a regression model where the target variable is binary, that is, it can take only two values, 0 or 1. It is the most utilized regression model in readmission prediction, given that the output is modeled as readmitted (1) or not readmitted (0). BLR is a statistical tool that classifies local node behavior to either malicious or benign. BLR has two stages: training and evaluation. At the training stage it uses node behavior from both benign and malicious node activity and derives a detection module. At the evaluation phase, data that was not used in the training stage, is used to evaluate the detection model. Binary Logistic Regression has the following assumptions:

- adequate sample size
- absence of multicollinearity
- no outliers

According to Fernandez Delgado et al. (2014) [10] "The classifier most likely to be the best are the random forest versions, the best of which (implemented in R and accessed via caret), achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets." This quote clearly pointed the powerful of RF in classification filed. The basic idea of RF is to produce numerous trees and combine the results. The random forest technique does this by applying two different tricks in model development. The first is the use of bootstrap aggregation or bagging, as it is called.

In the bagging process, a single tree is built on a random sample of the dataset, which accounts for about two-thirds of the total observations (note that the remaining third is called out-bag (oob)) . Repeat this dozens or hundreds of times, and then calculate the average of the results. The growth and pruning of each tree is not based on any error measure, which means that the variance of each tree is high. However, by averaging the results, you can reduce the variance without increasing the bias. The next thing that random forest brings to the table is that concurrently with the random sample of the data, that is, bagging, it also takes a random sample of the input features at each split. We will use the default random number of the predictors that are sampled, which, for classification problems, is the square root of the total predictors.

The advantage of RF by doing this random sample of the features at each split and incorporating it into the methodology, you can mitigate the effect of a highly correlated predictor becoming the main driver in all of your bootstrapped trees, preventing you from reducing the variance that you hoped to achieve with bagging. The subsequent averaging of the trees

that are less correlated to each other is more generalizable and robust to outliers than if you only performed bagging.

5. Results

6. References

[1] R. M. Krauss, V. Geller, and C. Olson, "Modalities and cues in the detection of deception," in *84th Annual Convention of the American Psychological Association, Washington, DC*, 1976.

[2] S. Venkatesh, R. Ramachandra, and P. Bours, "Robust algorithm for multimodal deception detection," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 534–537.

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv:1107.4557*, 2011.

[4] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 29–38.

[5] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[6] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[7] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature extraction*. Springer, 2006, pp. 315–324.

[8] T. Xiong and V. Cherkassky, "A combined svm and lda approach for classification," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 3. IEEE, 2005, pp. 1455–1459.

[9] K. Hugenberg, A. R. McConnell, J. W. Kunstman, E. P. Lloyd, J. C. Deska, and B. Humphrey, "Miami university deception detection database," 2017.

[10] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.