**The template is designed to help your submission, for more detailed information about *Scientific Reports* and its submission guidelines, please go to the journal’s official website.**

* Homepage: [nature.com/srep/](https://www.nature.com/srep/?utm_source=other_website&utm_medium=display&utm_content=mpu&utm_campaign=JRCN_3_CZ01_CN_SciRep_Template)
* Submission guideline: [nature.com/srep/author-instructions/submission-guidelines](https://www.nature.com/srep/author-instructions/submission-guidelines?utm_source=other_website&utm_medium=display&utm_content=mpu&utm_campaign=JRCN_3_CZ01_CN_SciRep_Template)
* Submit link: [author-welcome.nature.com/](https://author-welcome.nature.com/)

**Submission notes:**

* No strict word limits are imposed, but it is suggested that the main text (excluding the abstract, methods, reference, and figure legends) does not exceed 4500 words. Additionally, the submission should be no more than 11 typeset pages in length.
* Figures may be inserted within the text at the appropriate positions or grouped at the end.
* Equations must be inserted as editable equations (i.e.,) and not as images and identified by parenthetical numbers, such as (1), and should be referred to as "equation (1)" in the text.
* For the convenience of the peer review process, line number could be added in the manuscript.

**Deception Detection using Random Forest-based Ensemble Learning**

Kun Bu1,\*, and Kandethody Ramachandran1

1University of South Florida, Department of Mathematics & Statistics, Tampa, FL 33620-5700, USA  
\*kunbu@usf.edu

Abstract

The purpose of this work is to detect people lying using different ensemble machine learning algorithms to conclude a better classification model through comparison. Random Forest (RF) did an efficient work while dealing with both classification and regression problems; In this paper, we proposed a Random Forest-based ensemble learning, which is the combination of RF with SVM, GLM, KNNs, and GBM to improve the model performance. The data set 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 data set into Random Forest-based ensemble learning models, which includes RF+SVM.Linear, RF+SVM.Poly, RF+GLM, RF+KNNs, RF+GBM (Stochastic Gradient Boosting) and RF+WSRF (Weighted Subspace Random Forest). As a comprehensive comparison of the model performance, we conclude our new combination of algorithms performs better than the traditional machine learning models. Our contribution in this work provides a robust classification method which improves the predicted performance while avoiding model overfitting.

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 a person is lying. In other words, the polygraph test is easy to pass for those well-trained people (i.e.. 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. 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. Our motivation also comes from those previous studies and aims to develop a better deception detection via machine learning tools.

Ensemble learning, sometimes referred to as a multi-classifier system, builds and combines multiple classifiers to complete the learning task. There are two choices for getting multiple classifiers. The first is supposed that all individual classifiers are of the same type, or homogenous. For example, both decision tree individual classifiers, or both neural network individual classifiers (i.e., Bagging and boosting, for example, Random Forest). The second is supposed that all individual classifiers are not homogeneous, or heterogeneous. For example, in this paper, we have a classification problem of deception detection, we use support vector machine (SVM) individual learner, logistic regression (LR) individual learner and k-Nearest Neighbors (KNNs) individual learner to learn the training set, and then determine the final strong classifier by some combination strategy. This integration is called Stacking. In the experimental section, we applied both Bagging, boosting and stacking, and selected a better ensemble model to classify people lying.

Methods

Exploratory Data Analysis

Data Description

MU3D is collected by recording 80 targets speaking honestly and dishonestly about their social relationships. The dataset was divided 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 12 variables with 1 label variable called Veracity in the video level dataset, a short variable description is shown in Table 1.

Statistical Data Analysis

Due to the data properties, normalizing the data so that they are of the same order of magnitude is much better for machine learning. Except for VideoID and the last variable Transcription, a normalized box-plot was created for all other variables by Veracity. As shown in Figure 1, only the variables Accuracy, TruthProp and Trustworthy have differences 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. Based on the variable’s linear correlation relationship in Figure 2, we can reduce the features dimension from 11 to 9, where the VidLength\_ms and TruthProp were removed because they are linearly correlated with VidLength\_sec and Accuracy, respectively.

Ensemble Learning for Deception Detection

Algorithms Selecting Procedures

The next step is to fit MU3D into a machine learning model and see how the computer performs on detecting lies. We first trained three different models based on the data properties, including Support vector machine (SVM), Binary Logistic Regression (BLR) and Random Forest (RF) to predict the deception. The basic idea of these three selected algorithms is based on the Figure 3 algorithm flowchart. The purpose of this algorithm flowchart is to create a tool that helps not only select the possible modeling techniques but also understand deeper about the problem itself. As shown in Figure 3, by answering the following questions from the flowchart, we initially decided the main three modeling techniques to be applied in this prediction. In the next subsection, we explained the reason why these three machine learning algorithms were chosen based on the data properties and the model assumptions.

Preliminaries Machine Learning Algorithms

Support Vector Machines are based on a decision plane concept that defines decision boundaries, and SVMs have been shown to perform well in a variety of settings, and are often considered one of the best ”out of the box” classifiers according to James, G. et al.(2013) James et al. (2013). 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. According to the statistical data analysis section, the data properties show that 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 excepted the Transcription are 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 tens, hundreds and thousands of features. In our dataset, we have more than 10 features which motivates us to choose SVM. Another reason 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: 1) The distance from the SVM’s classification boundary to the nearest data point should be as large as possible; the distance formulas include Euclidean distance, Manhattan distance, Chebyshev distance and Minkowski distance. Where the Euclidean distance and Manhattan distance are special forms of the Minkowski distance, and 2) The support vectors are the most useful data points because those points are the ones most likely to be incorrectly classified. This means that the primary goal of training SVMs is to find support vectors in the dataset that both separate the data and find the maximum margin between classes.

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 deception prediction, given that the output is modeled as truth (1) or lie (0). BLR is a statistical tool that classifies the MU3D target person in a video to either lie or not. BLR has two stages: training and evaluation. At the training stage it uses Video level data from both lie and truth and builds a detection module. At the evaluation stage, data that was not used in the training stage, is used to evaluate the detection model.

Logistic regression is also called logit (log unit) regression, and we usually build a Logistic regression model by starting to build the model by combining the generalized linear model with logit function, or from the point of a random variable following the logistic distribution. Binary Logistic Regression has the following assumptions: 1) adequate sample size, 2) absence of multicollinearity and 3) no outliers. Note that according to EDA, the outliers in our dataset need to be removed before fitting this model so that it doesn’t violate the assumptions.

The mathematics expression of the logistic regression is

Where is called logit function and it is the link function for the generalized linear model. This logistic regression model can be expressed as

Where is the natural logarithm, and the above function is called sigmoid function.

According to Fern ́andez-Delgado et al. (2014), ”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 out the power of RF in the classification field. 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 in short as bagging. In the bagging process, a single decision 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-of-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 according to Merentitis et. al (2014).

Results

The results section can include subheadings.

Discussion

The discussion must not contain subheadings.

Data availability

*Scientific Reports* requires the inclusion of a data availability statement with all submitted manuscripts, as this journal requires authors to make available materials, data, and associated protocols to readers.

References

1.

2.

E.g. Author 1, Author 2, Author 3 & Author 4. Title. *Journal name abbreviation*, **volume number**, page number (Year).

Figure legends (provided in numerical order)

Figure legends are limited to 350 words per figure. They should begin with a brief title sentence and include a short description of each panel and the symbols used (i.e., ▲— control group; ▲—treatment group 1; ▲—treatment group 2).

Tables

**Table 1.** Tables are submitted in the main article file in an editable format (see below) and not as images. Tables containing statistical data analysis require descriptions of standards of error analysis and ranges in the table caption (e.g., n=3, SD=standard deviation).

|  |  |  |
| --- | --- | --- |
| Parameter | A (±SD) | B (±SD) |
| X | 1.01±.02 | 2.02±01 |

Competing interests (mandatory)

A competing interests statement must be given on behalf of all authors. Competing financial and non-financial interest must be disclosed. If there are no competing interests, a statement must still be given (e.g., the authors declare no competing interests).

Acknowledgements (optional)

Author contributions

Names must be given as initials. For example: B.W. conceived the experiment(s). R.W. and D.W. conducted the experiment(s), and B.W. performed statistical analysis and figure generation. All authors reviewed the manuscript.