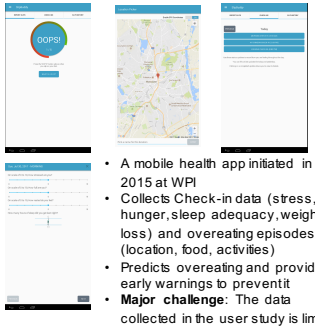
Using Synthetic Data to Enhance Multiple Machine Learning Techniques in Mobile Health Applications

Khang Vinh Tran [1]
Advisor: Carolina Ruiz [2], Co-Researcher: Katherine Finnerty [3]
[1] UC Berkeley/De Anza College , [2] Worcester Polytechnic Institute, [3] Amherst College

Abstract

SlipBuddy is a mobile health application that tracks overeating behavior, and uses machine learning to predict and prevent overeating. A user study was conducted to evaluate SlipBuddy's efficacy. However, user studies are limited in terms of number of participants and data collected from each participant. This project investigated using synthetic data to enhance predictive performance of the real data collected in the user study. The process started with generating synthetic data that follow the same distribution of the original observed data and preserve the correlation between predictors and overeating behavior. Then machine learning techniques were applied to three separate datasets: the original data, the synthetic data, and the original and synthetic data combined. The predictive performance of the resulting machine learning models were compared, and the statistical significance of the results evaluated. Future work will investigate alternative synthetic data generation methods, and new ways of leveraging synthetic data in machine learning.

SlipBuddy



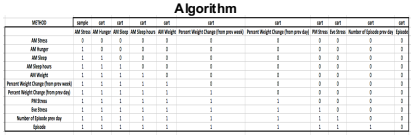
- A mobile health app initiated in 2015 at WPI
- Collects Check-in data (stress, hunger, sleep adequacy, weight loss) and overeating episodes (location, food, activities)
- Predicts overeating and provides early warnings to prevent it
- **Major challenge**: The data collected in the user study is limited in size

Methodology

- Generate synthetic data
- Construct predictive model with different machine learning techniques using k-fold cross validation
- Evaluate and compare the predictive performance of different machine learning models using multiple performance metrics
- Platforms: IPython Notebook and R Studio

Synthetic Data Generation

Algorithm

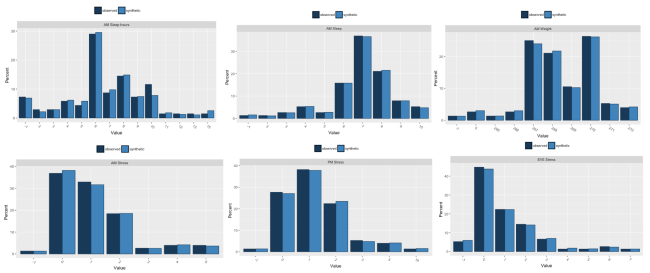


Using synthpop packages on R [2]:

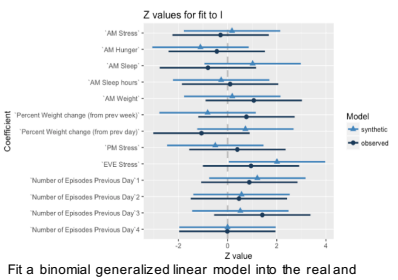
- Fit the observed data into a sequence of regressions
- First variable is randomly sampled with replacement from real data
- The values of each subsequent variable are drawn from a posterior distribution yielded by the cart regression of the previous variables

Two key properties of the original data that must be maintained:

- The probability distribution of data within each variable
- The correlation to Overeating Episode of each variables

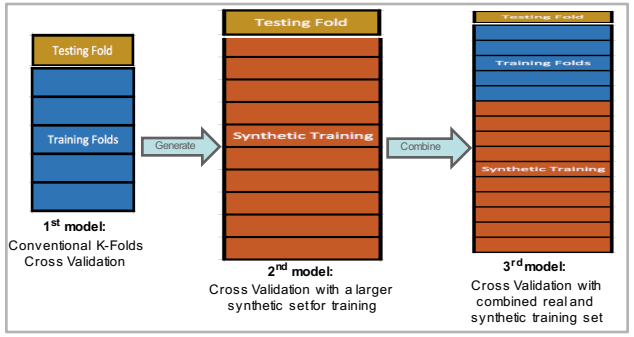


Histograms showing the similarity of relative frequency distribution between real data and synthetic data



Fit a binomial generalized linear model into the real and synthetic sets to measure the consistency of correlation from each variable and Overeating Episodes

Comparing Decision Trees and K-Nearest Neighbors on Real and Synthetic Data



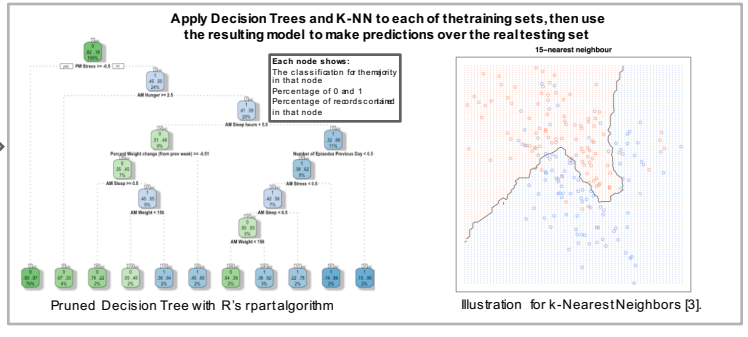
1st model: Conventional K-Folds Cross Validation

2nd model: Cross Validation with a larger synthetic set for training

3rd model: Cross Validation with combined real and synthetic training set

For each data set of CV
Train and Test

Apply Decision Trees and K-NN to each of the training sets, then use the resulting model to make predictions over the real testing set



Pruned Decision Tree with R's rpart algorithm

Illustration for k-Nearest Neighbors [3].

Performance Comparison Result and Conclusion

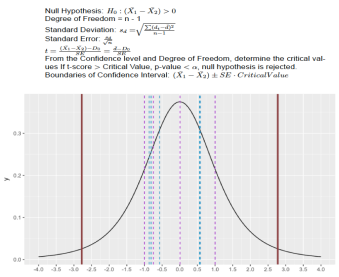
Performance metrics to evaluate and compare models:

- Accuracy
- Precision
- Recall
- Fallout
- F-Measure
- AUC

Paired t-test for the mean of each of the metrics in:

- Real Training vs. Synthetic Training
- Real Training vs. Real-Synth combined
- Synthetic Training vs. Real-Synth combined

*** Repeat this process for both Decision Tree and KNN results**



T distribution, degree of freedom = 4, Confidence Level: 95%,
brown line: Critical value, blue dashes: Tree measures' t score, pink dashes: KNN's Measures' t score
As shown on the plot, none of the t statistics from either two machine learning techniques can exceed the critical value

Conclusion:

- No statistical significance to conclude that synthetic data enhanced the predictive performance for all models (at the 95% confidence level, p-value < 0.05)
- For some users, K-Nearest Neighbors give better accuracy as well as AUC than Decision Trees

Future Work:

- Investigate the reason why synthetic data did not outperform the limited real data. With the variable correlation variation as a possible cause
- Apply different Machine Learning techniques (e.g., Naïve Bayes, Random Forest), and compare the results with existing results

References

[1] Adult obesity in the United States [online]. Available: https://stateofobesity.org/adult-obesity/
[2] Nowok Beata, Gillian M. Raab, and Chris Dibben. "Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R." Statistical Journal of the IAOS Preprint (2015): 1-12.
[3] T Hastie, R Tibshirani, J Friedman. *Element of Statistical Learning*. Spinger 2009.

Acknowledgement

This research was funded by the National Science Foundation (NSF) through the **REU SITE: Data science research for safe, sustainable and healthy communities (CNS – 1560229)**
This research was also conducted with the support and insightful feedback of Cole Polychronis, Brendan Foley, Quyen Hoang, Charan Sankaran, Vanshika Chowdhary, and Professor Begisu Tulu from WPI.