Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# Extreme Gradient Boosting and Behavioral Biometrics

**Benjamin Manning**

University of Georgia, Athens, GA 30602 USA (benjamin.manning@uga.edu)

**Abstract**

As insider hacks become more prevalent it is becoming more useful to identify valid users from the inside of a system rather than from the usual external entry points where exploits are used to gain entry. One of the main goals of this study was to ascertain how well Gradient Boosting could be used for prediction or, in this case, classification or identification of a specific user through the learning of HCI-based behavioral biometrics. If applicable, this procedure could be used to verify users after they have gained entry into a protected system using data that is as human-centric as other biometrics, but less invasive. For this study an Extreme Gradient Boosting algorithm was used for training and testing on a dataset containing keystroke dynamics information. This specific algorithm was chosen because the majority of current research utilizes mainstream methods such as KNN and SVM and the hypothesis of this study was centered on the potential applicability of ensemble related decision or model trees. The final predictive model produced an accuracy of 0.941 with a Kappa value of 0.942 demonstrating that HCI-based behavioral biometrics in the form of keystroke dynamics can be used to identify the users of a system.

## Trees, Ensembles and Gradient Boosting

Decision trees are widely used in classification type problems of this nature as they can predict the value of a dependent variable just as many other algorithms do, by learning the values contained in the features of the data or the independent variables. In this approach the highest correlating features are used to eventually categorize or classify related classes of the target or dependent variable. In this recursive process (Quinlan 1986) 'splits' or 'branches' of the trees are constructed from each correlated feature of the data until the correlations are no longer effective or the target variable is reached at the bottom of the tree. Along these individual paths binary splits are derived and are presented often as binary decisions, which outline the overall relationship the dependent variable has with the remaining features in the data set.



One main disadvantage of decision trees is the high probability of becoming unstable when used on data with larger numbers of features. In order to remedy this potential side effect ensemble methods (Dietterrich 2000) can be used to introduce an iterative training process.

This iteration uses different methods that invoke building and comparing the results of numerous models during a single model building session. Random Forest is a good example of an ensemble learning method where multiple decisions trees (i.e. the forests) are created during training and the output is derived by taking the mean of all of the produced trees within the forest, but a less often used and often overlooked ensemble approach is Gradient Boosting.

Gradient Boosting (Friedman 2002) constructs additive models by "sequentially fitting a simple parameterized function to current residuals by using the least square at each iteration." Put more simply, the model builds multiple decision trees similar to how Random Forest does, but Gradient Boosting uses an arbitrary differential loss function instead of using the mean to make the resulting predictions.

## Data Science Process

### Data Collection

The keystroke dynamics data (Killourhy and Maxion 2009) for this study was recorded from fifty-one typists typing the same single word (.tie5Roanl) four hundred times. The researchers constructed a data collection system that recorded different keystroke events such as key-down and key-up as well as the name of the key that had been pressed and the correlating time between different key events. If participants made any errors when entering the data, they were instructed to retype the word again and continue with the remaining iterations. This data was then analyzed to create a word-timing table dataset that contained 20,400 individual observations and thirty-four variables: thirty-three features represented the timestamp of the keystrokes involved in typing the single word and one variable, representing the dependent variable,

represented the identity of the person doing the related typing tasks.

### Preprocessing, Model Description and Results

The dataset was analyzed to ascertain if any of the features could be removed from dataset. Two features: one representing the individual session identification and a second feature representing the individual observation number, were removed from the data to ensure that no bias was created by features associated with identification only. This reduced the dimensionality of the dataset to thirty-one features and one dependent variable; these elements were used to train and test the model for the study.

The Classification and Regression Training (caret) package for R was chosen for model training and testing and the dataset was randomly stratified and divided into separate training and testing sets containing 70% of the data and 30% of the data, respectively.

The algorithm chosen for this study, Extreme Gradient Boosting (XGB), is a combination of Gradient Descent (Burges, et al. 2005) and Boosting (Dietterich 2000) and offers different tuning parameters that were modified for with the goal of building an optimal model. The tuning parameters included number of iterations, maximum tree depth, shrinkage, minimum loss reduction, subsample rates and minimum sum of instance weights. In order to create a necessary comparison with XGB models were also created using C50 and KNN so the benefit of using XGB could be easily seen. Similar tuning parameters for both of the additional algorithms were also established before training began.

When using XGB the number of iterations specified how many times the data would be analyzed; 150 was chosen as the optimal number of iterations through numerous cycles of training to reduce any unnecessary training time. The maximum depth of the model tree was limited to just two branches to adequately prevent overfitting. Shrinkage was set at .3 to ensure the model would be strong enough to generalize to new data when making predictions while also improving the performance of the model in the most optimal way. The minimum loss reduction or Gamma was set at 0 for the entire training process and the fraction of observations selected (subsample rates) for each tree was set at 0.6.

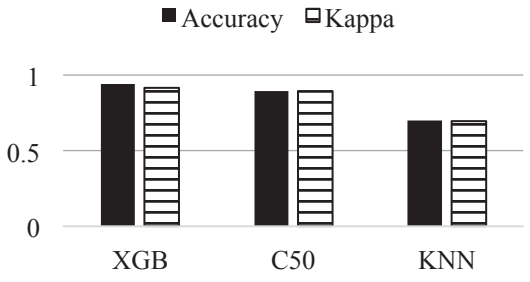


*Table I – Model Metric Comparison*

Lastly, the minimum sum of instance weights default setting of 1 was left unchanged for training. Using the parameters previously described, the models produced results as shown in Table I.

## Conclusion

These results are favorable for the use of keyboard dynamics data to identify that the user of system is the same user the training data was obtained from. This process could also be scaled and expanded to allow for extended training using larger amounts of unstructured data such as daily postings in social media or common daily and routine tasks that require text entry into the system.

Typing is an integral input interaction for a vast majority of user based software systems and it would be easy to integrate a similar predictive model into the lifecycle of software development for high security scenarios. Minimally, this would help to prevent inside attacks where users that are not the owners of valid credentials try to enter passwords and other login information that are not their own, but are deemed valid for entry into the system. In addition, this prospective model could also be used to assist verification during larger multi-factor authentication scenarios where users enter simple username/password combinations along with other text information - such as the answers to challenge questions.

Both of these scenarios provide two examples, one for each part of a system, that demonstrate how Gradient Boosting models and machine learning can be used to prevent attacks from the inside of a system by identifying users through the validation of an identity using behavioral biometrics in the form of keyboard stroke dynamics; this is notable as the findings of this study are part of a much larger proposal on designing a decision support system for multifactor authentication.

## References

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005, August). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning* (pp. 89-96). ACM.

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer Berlin Heidelberg.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378.

Killourhy, K. S., & Maxion, R. A. (2009, June). Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks* (pp. 125-134). IEEE.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.