User Authentication through Keystroke Dynamics

Mid Stage Report

CS725 Project

Indian Institute of Technology, Bombay
Department of Computer Science and Engineering

Devashish Singh (163059001)
Prateek Patidar (163059006)
Shubham Singh (163059008)
Hareesh Kumar (16305R013)

1 Project Description

Keystroke dynamics—the analysis of typing rhythms to discriminate among users—has been proposed for detecting impostors (i.e., both insiders and external attackers). Since many anomaly-detection algorithms have been proposed for this task, it is natural to ask which are the top performers (e.g., to identify promising research directions). Unfortunately, we cannot conduct a sound comparison of detectors using the results in the literature because evaluation conditions are inconsistent across studies.

2 Dataset Collection

As the subject types the password, it is checked for correctness. If the subject makes a typographical error, the application prompts the subject to retype the password. Whenever the subject presses or releases a key, the event (i.e., keydown or keyup), along the key involved, and a timestamp for the moment at which the keystroke event occurred. An external reference clock was used to generate highly accurate timestamps. The reference clock was demonstrated to be accurate to within 200 microseconds (by using a function generator to simulate key presses at fixed intervals). 51 subjects (typists) typed the same password, and each subject typed the password 400 times over 8 sessions (50 repetitions per session). They waited at least one day between sessions, to capture some of the day-to-day variation of each subject's typing. The password (.tie5Roanl) was chosen to be representative of a strong 10-character password.

3 Dataset Information

The data are arranged as a table with 34 columns. Each row of data corresponds to the timing information for a single repetition of the password by a single subject. The first column, subject, is a unique identifier for each subject (e.g., s002 or s057). The second column, sessionIndex, is the session in which the password was typed (ranging from 1 to 8). The third column, rep, is the repetition of the password within the session (ranging from 1 to 50). The remaining 31 columns present the timing information for the password. The name of the column encodes the type of timing information. Column names of the form H.key designate a hold time for the named key (i.e., the time from when key was pressed to when it was released). Column names of the form DD.key1.key2 designate a keydown-keydown time for the named digraph (i.e., the time from when key1 was pressed to when key2 was pressed). Column names of the form UD.key1.key2 designate a keyup-keydown time for the named digraph (i.e., the time from when key1 was released to when key2 was pressed). Note that UD times can be negative, and that H times and UD times add up to DD times.

4 Papers

- Comparing Anomaly Detectors for Keystroke Dynamics[1]
- ROCR: visualizing classifier performance in R [2]

5 Approach

We did some feature engineering on the dataset i.e. added square and square root of features. We then used some classification methods available in scikit-learn library on the data like Logistic Regression, Support Vector Machines, Random Forests, K NN Classification,MLP and Gaussian Naive Bayes.

6 Results

Results without Feature Engineering:
Logistic Regression : 0.702777777778
Support Vector Machines : 0.0970588235294
Random Forests : 0.873366013072
K NN Classification : 0.157679738562
Gaussian Naive Bayes : 0.57091503268

Results after Feature Engineering:
Logistic Regression : 0.776307189542
Support Vector Machines : 0.0197712418301
Random Forests : 0.854738562092
K NN Classification : 0.15637254902
Gaussian Naive Bayes : 0.37385620915

The best achieved accuracy is about 0.85 by using Random Forest.

7 Work to be done

- Using online approach instead of batch.
- Feature engineering using trigrams and other methods.
- Tuning the parameters of the mentioned classification algorithms so as to achieve more or less same accuracy as achieved in the paper.
- Using another approach other than classification that is of correlation using basic techniques and siamese networks.

References

[1] Kevin S. Killourhy and Roy A. Maxion. Comparing anomaly detectors for keystroke dynamics. In *Proceedings of the 39th Annual International Conference on Dependable Systems and Networks*, DSN-2009, pages 125–134, New York, NY, USA, 2009. IEEE.

[2] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940, 2005.