Biometric Authentication of Smartphone Users with Support Vector Machines

Nico Espinosa Dice

April 6, 2020

Presentation Outline

- The Security-Convenience Trade-off
- Background on Biometric Authentication
- Dataset: Biometric Data from Mobile Devices
  - Exploratory Data Analysis
- **Hypothesis**
- **Mathematics of Support Vector Machine Models**
  - **Cross-Validation and Hyperparameter Tuning**
  - **Mathematical Backing for Selected Hyperparameters**
- **Authentication Results and Model Comparison**
- Discussion and Future Research

The Security-Convenience Trade-off

- **Security**: Essential for the safety of users and companies.
  - Security failures pose existential threats to reputation-oriented companies.
- **Convenience**: Crucial for maintaining a high yield of returning users.
  - Lack of convenience drives users to competitors.
- *General Trend*: High security is less convenient.

Background on Biometric Authentication

- **Biometrics**: Metrics (data) acquired from human characteristics.
  - Body measurements and human behavior.
- **Biometric Authentication**: Collect biometric data and use machine learning techniques to authenticate (distinguish) users.
- *Advantages*:
  - Requires little effort from users.
  - High security with convenience.

Dataset: Biometric Data from Mobile Devices

- Provided by: IDSeal, a cybersecurity company.
- Year: 2014.
- Type: Acceleration data.
- Source: Inertial measurement units of smartphones.
- Data collected during "normal device usage" over a period of several months.

Exploratory Data Analysis

Features of Data

- T = time (milliseconds).
- X = acceleration (g) in x direction.
- Y = acceleration (g) in y direction.
- Z = acceleration (g) in z direction.
- DeviceId (Training) = Unique ID of device.
- SequenceId (Testing) = Unique number assigned to each test sample.

Data Shape

- Total Samples: 60,000,000.
- Total Devices: 387.
- Test Data: 90,000 consecutive samples per device.
- Every device had more than 6000 samples.
- Zero-movement periods lasting 10+ seconds were removed.

Support Vector Machine

Chosen because of recommendations in literature (Sitova et al. 2016).

Equation 14.59 of Murphy 2012

$$\hat{y}(\boldsymbol{x}) = sgn\left(\hat{w}_0 + \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})\right), \text{ where}$$

$$\boldsymbol{x}_i : \text{support vector (when } \alpha_i > 0),$$

$$\alpha_i = \lambda_i y_i,$$

$$k : \text{kernel function.}$$

Kernelized SVMs: $O(n_{\text{features}} \times n^2_{\text{observations}})$ complexity (Murphy 2012; Sitova et al. 2016).

Support Vector Machine: Kernel Functions

Kernel: Linear Function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$$

Kernel: Polynomial Function

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}') &= (\frac{\boldsymbol{x}^\top \boldsymbol{x}'}{2\sigma^2} + c_0)^d \\ &= (\gamma \boldsymbol{x}^\top \boldsymbol{x}' + c_0)^d, \text{ where} \\ \gamma &= \frac{1}{2\sigma^2}, \\ \sigma^2 &: \text{“bandwidth,”} \\ d &: \text{kernel degree.} \end{aligned}$$

Support Vector Machine: Kernel Functions

Kernel: Sigmoid Function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\gamma \boldsymbol{x}^\top \boldsymbol{x}' + c_0), \text{ where}$$
$$\gamma : \text{"slope,"}$$
$$c_0 : \text{"intercept."}$$

Kernel: Radial Basis Function

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}') &= exp(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2\sigma^2}) \\ &= exp(-\gamma ||\boldsymbol{x} - \boldsymbol{x}'||^2), \text{ where} \end{aligned}$$
$$\gamma = \frac{1}{2\sigma^2},$$
$$\sigma^2 : \text{"bandwidth."}$$

Support Vector Machine: RBF Kernel

Advantages:

- Outliers have less impact.
- Effective in higher dimensions.
  - Important when adding more features: gyroscope and magnetometer data.
- **Excels when intersection of classes is trivial**.
  - Biometric authentication: no overlap between classes.

Disadvantages:

- Long fitting time.
  - Poses difficulties when cross-validating.
- Difficult to visualize and interpret.

Hypothesis

RBF kernel outperforms other kernel functions, specifically:

–Linear,

–Polynomial, and

–Sigmoid.

Comparing Kernels

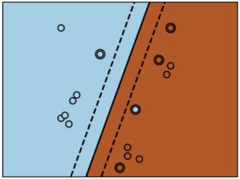
Figure 1: Linear Kernel

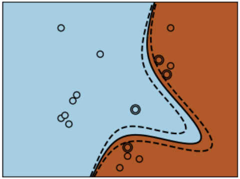
Figure 2: Polynomial Kernel

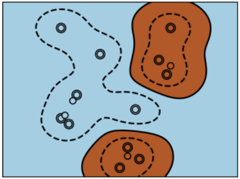
Figure 3: RBF Kernel

Source: Scikit-learn.org

Comparing Kernels: Results

Kernel Function	C	γ	**Degree**	**Accuracy**
Sigmoid	1	1/(Number of Features)	N/A	53%
Linear	1	N/A	N/A	70%
Polynomial	1	1/(Number of Features)	3	77%
RBF	1	1/(Number of Features)	N/A	79%

Table: Accuracies of SVMs with Varying Kernel Functions

All SVMs were implemented using the recommended values for C and γ that were found in the literature.

Cross-Validation

Cross-Validation for Kernel Function and C

2-dimensional grid consisting of values:

$k \in \{\text{Linear, Polynomial, Sigmoid, RBF}\}$
$C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$

Cross-validation range for C empirically determined in literature (Hsu et al. 2009; Murphy 2012).

Results:

- Kernel : Radial Basis Function
- $C = 1$

Hyperparameter Selection

Hyperparameters of Support Vector Machine

γ (Gamma) $= \frac{1}{\text{Number of Features} \times \text{Variance of Z-Acceleration}}$
Kernel = Radial Basis Function (RBF)
C (Regularization Parameter) = 1
Tolerance for Stopping Criterion $= 1 \cdot 10^{-3}$
Shrinking Heuristic = True

Cross-Validation for C and γ

2-dimensional grid consisting of values:

$C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$
$\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$

Cross-validation range empirically determined in literature (Hsu et al. 2009; Murphy 2012).

Cross-Validation cont.

Cross-Validation for C and γ

2-dimensional grid consisting of values:

$C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$
$\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$

Implementation:

- Randomized search through 2-D grid of values
  - Did not test every combination of values

Results:

- $C = 2^9 = 512$,
- $\gamma = 2$.

Authentication Results: Midterm Project

Because of the size of the dataset, model was trained and tested on subset of the dataset: 5 devices.

- Model produced 79% accuracy without cross-validating hyperparameters.
- Accuracy:
  - For each test sample: predicted its corresponding device.
  - 'Accuracy' is the percentage of correct predictions.

```
In [85]: svm2.getResults()
Out[85]: 'Accuracy: 0.7904519072701044'
```

Figure 4: Accuracy of support vector machine using radial basis function kernel.

Authentication Results: Final Project

Because of the size of the dataset, model was trained and tested on subset of the dataset: 5 devices.

Kernel Function	C	γ	**Degree**	**Accuracy**
Sigmoid	1	1/(Number of Features)	N/A	53%
Linear	1	N/A	N/A	70%
Polynomial	1	1/(Number of Features)	3	77%
RBF	1	1/(Number of Features)	N/A	79%
RBF (CV)	2^9	2	N/A	83%

Table: Accuracies of SVMs with Varying Hyperparameters

Tradeoff: Accuracy vs. Number of Devices

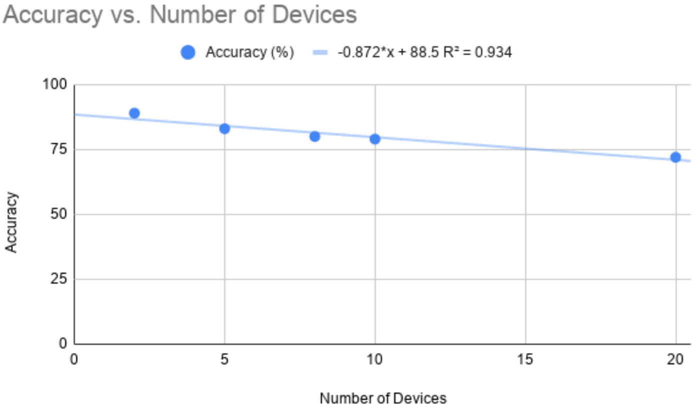

Figure 5: The accuracy of an SVM trained and tested on varying numbers of devices.

Discussion and Future Research

Discussion:

SVM shows potential for biometric authentication.

- Results support hypothesis: RBF outperforms other kernel functions.
- Empirically determined C, γ values differ from literature.

Future Research:

- Analyze trade-off between amount of training data (runtime) and accuracy.

Sources

- Bhattacharyya, Debnath Ranjan, Rahul Alisherov, Farkhod Minkyu, Choi. (2009). Biometric Authentication: A Review. International Journal of u- and e- Service, Science and Technology. 2.
- Murphy, Kevin P. Machine Learning: A Probabilistic Perspective. 4th ed., Cambridge, MIT Press, 2013.
- R., Malathi Jeberson Retna Raj, Retna. (2016). An Integrated Approach of Physical Biometric Authentication System. Procedia Computer Science. 85. 820-826. 10.1016/j.procs.2016.05.271.
- Sitová, Zdeňka Sedenka, Jaroslav Yang, Qing Peng, Ge Zhou, Gang Gasti, Paolo Balagani, Kiran. (2015). HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users. IEEE Transactions on Information Forensics and Security. 11. 1-1. 10.1109/TIFS.2015.2506542.
- Wang, Jindong Chen, Yiqiang Hao, Shuji Peng, Xiaohui Lisha, Hu. (2017). Deep Learning for Sensor-based Activity Recognition: A Survey. Pattern Recognition Letters. 10.1016/j.patrec.2018.02.010.