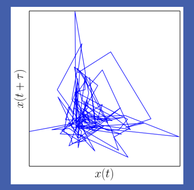
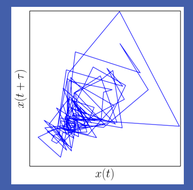


Classification and Authentication of One-dimensional Behavioral Biometrics

John V. Monaco

Pace University, Pleasantville, NY 10570, USA



What is a one-dimensional behavioral biometric?

A one-dimensional time series that contains values observed as a result of some aspect of human behavior.

For example,

- Hitting a single key on a keyboard
- Timestamps from encrypted network traffic
- Anonymized web history timestamps
- Horizontal eye movement
- Telephone call log timestamps
- Bitcoin transaction timestamps

Do you see a pattern? One-dimensional behavioral biometrics frequently arise in situations where the information is encrypted or anonymous. But, this doesn't necessarily deter user identification.

Methodology

A proper embedding is first found for each dataset. The embedded samples are compared to each other using the **Wald-Wolfowitz** test. Classification and authentication results are obtained using the WW statistic as a distance measure in a **KNN classifier**.

Classification and authentication results are obtained for several publicly available datasets. All datasets are trimmed to approximately 60 users and contain exactly 7 sessions per user and 130 events per session. Users and sessions are selected randomly.

In all of the following datasets, only the event timestamps are used

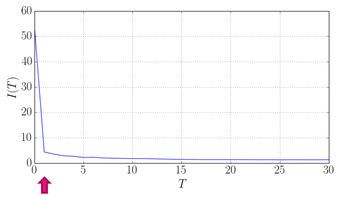
Dataset	Description	Source
Password	Users entering the password .tie5Roanl	CMU keystroke: http://www.cs.cmu.edu/~keystroke/
Mouse	Undergraduate students taking online exams	Author: http://vmonaco.com/datasets
Keystroke	Free-text keystrokes from students answering open-ended questions	Author: http://vmonaco.com/datasets
Key-blow	Users repeatedly hitting a single key on a keyboard	2014 RTI contest: http://www.upcv.upatras.gr/rti/
Web history	Anonymized web history donated by participating users	Web History Repository: http://webhistoryproject.blogspot.com/

Embedding Procedure

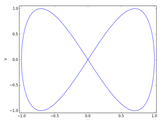
1. Determine uniform embedding parameters, d_e and τ

τ is found first using the mutual information, where $I(T)/I(0) \approx 1/5$

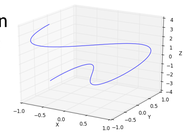
This gives us **τ=1**



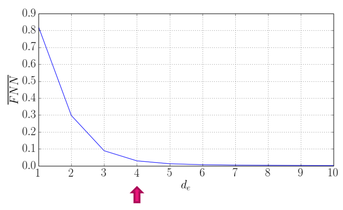
The method of **false nearest neighbors** (FNN) is used to determine the embedding dimension.



Two points are FNN when they are no longer close in a higher dimension



At d_e = 4, less than 5% of the embedded vectors are FNN



2. Define an embedding window $d_w = d_e \times \tau$ and search for the optimal lag vector using the minimum description length (MDL) principle as a heuristic. The description length of the embedded time series is minimized:

$$DL(\mathbf{X}) = \frac{d}{2}\ln\left[\frac{1}{d}\sum_{i=i}^{d}(y_i - \bar{y})^2\right] + d + DL(d) + \frac{n-d}{2}\ln\left[\frac{1}{n-d}\sum_{i=d+1}^{n} e_i^2\right]$$

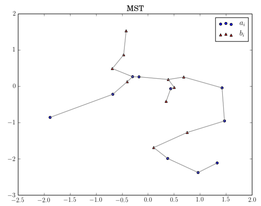
Integer description length

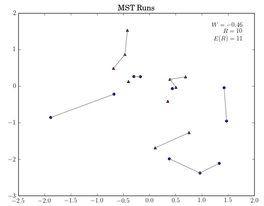
Asymptotically, this amounts to minimizing the model prediction error as $n \to \infty$

Approximate Wald-Wolfowitz Test

The multivariate Wald-Wolfowitz test is a non-parametric test to determine whether two samples come from the same distribution.

It relies on construction the minimum spanning tree (MST) among all observations in both samples and counting the number of runs. A **run** is a segment of the tree that touches vertices from the same sample.





Here is an MST with 10 runs: did these points come from the same distribution? (Yes the did)

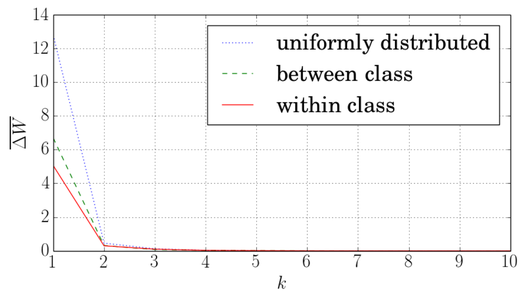
Expected number of runs: $E(R) = \frac{2mn}{N} + 1$

Wald-Wolfowitz statistic: $W = \frac{R - \frac{2mn}{N} - 1}{\left(\frac{2mn(2mn-N)}{N^2(N-1)}\right)^{\frac{1}{2}}}$

This operation is **O(N^3)**, which is not good for large datasets. Can we approximate the MST and still obtain reliable results?

Yes, consider only the **k-nearest neighbors** to each vertex when constructing the MST. This can be done efficiently with a k-d tree.

The runtime is reduced to **O(Nk log(Nk))**. Most data is well behaved, and the difference, ΔW, between the true W and approximate $W^{(k)}$ becomes negligible as k increases, where $W^{(k)}$ is the value found using the approximate MST:



Newell's Time Scale of Human Action

Where does each dataset lie on Newell's time scale?

Scale (sec)	Time Units	System	World (theory)
10^7	Months		SOCIAL BAND
10^6	Weeks		
10^5	Days		
10^4	Hours	Task	RATIONAL BAND
10^3	10 min	Task	
10^2	Minutes	Task	
10^1	10 sec	Unit task	COGNITIVE BAND
10^0	1 sec	Operations	
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	BIOLOGICAL BAND
10^{-3}	1 ms	Neuron	
10^{-4}	100 µs	Organelle	

Classification and Authentication Results

A KNN classifier is used to obtain results. ACC1 is the proportion of correctly classified samples, and EER is the point on the ROC curve at which FAR and FRR are equal. The event frequency and lag vector found for each dataset is also given.

Dataset	Freq. (Hz)	Embed.	ACC1(%)	EER(%)
Password	8.5	[1,2]	33.3	18.3
Mouse	9.0	[1,2,3]	34.5	17.1
Keystroke	4.2	[1,2]	43.1	12.7
Key-blow	3.4	[1,2]	44.1	14.0
Web	6.8×10^{-5}	[1,…,9]	16.2	32.8

Future Work

These results are promising, although very preliminary. As privacy is becoming an increasingly important issue, more research is needed to evaluate the potential of one-dimensional behavioral biometrics.

We should look at applications where much of the information is encrypted or inaccessible. For example, can a user be identified by a history of Bitcoin transactions, encrypted network traffic, or forum postings?