INTRODUCTION

While mobile devices are no longer a new technology, using the data generated from the use of these devices for security purposes has just been recently explored. Previous methods, such as passwords, are quickly becoming antiquated, lacking the robustness, accuracy, and convenience desired to serve as reliable security measures. Thus, researchers have resorted to alternative techniques; for instance, measurements obtained from keyboard interactions and user movement have been used to positively confirm the identity of the device’s owner. Similarly, behavioral interactions, such as application usage, have also been useful in maintaining secure use sessions on the device [1].

While several techniques have been applied for user identification, there lacks an extensive discussion regarding those used in smartphone usage data. This can further insight into mobile device usage data which can assist in improving identification performance, feature generation and selection, denoising, and data acquisition. This work explores these problems using data captured through smartphone interactions; specifically, association analysis is applied to a dataset of application, Bluetooth, and Wi-Fi usage from 189 users. The Apriori Algorithm is applied to various time-based partitions of the dataset for better understanding of how to extract frequency-based models which optimize user separability through the discovery of strong co-occurrences between actions and identities.

DATA & FEATURES

The dataset contains application, Bluetooth, and Wi-Fi usage data from 189 college-level students collected over 19 months. Each action in the dataset consists of a timestamp of data capture, action name (i.e., application, Bluetooth device, or Wi-Fi network name), and strength of the action (i.e., number of bytes received or transferred in application traffic or signal strength in Bluetooth and Wi-Fi).

Association rule mining discovers rules of the structure X → Y, where X and Y are termed the antecedent and consequent, respectively. A rule with X and Y as support and confidence percentages, respectively, is interpreted as X and Y each containing a subset of items, I, where items in X and Y appear in % of transactions, and % of transactions that contain X also contain Y. A transaction is a list of items which have been purchased together (in the market basket context), where all transactions are composed of items from I. Association rules discover frequent item sets, or items which occur frequently together, making them useful in discovering strong co-occurrences in large datasets.

Association rules are extracted from each data type separately and combined as behavioral features using 60% support and 90% confidence. To allow a continuous authentication scheme, features are extracted as follows:

1. The data stream is divided into week-long segments. This sampling rate is chosen based on preliminary research which indicated that these specific data types are most consistent over a week’s time [2].
2. For each week of data, actions performed by the user are sampled at 1, 5, 15, and 30 minute intervals. These actions are combined to form a transaction. These transactions are combined to form the set of transactions that are used for feature extraction.
3. For all association rules, only those in which the rule’s consequent is the user’s identity only are retained.

CONCLUSIONS

Association rule mining is employed to better understanding any underlying trends in mobile device usage data for the purpose of identification. This is a behavioral biometric approach, where a continuous authentication scheme is investigated. In practice, this could allow non-intrusive and transparent authentication for robust and convenient mobile device security. This work has been useful in evaluating several characteristics of the data.

Frequency-based representations are commonly used for identifying mobile device users in behavioral profiling, but there lacks an extensive analysis of this representation. The presented results suggest that frequency-based models are generally unique and consistent. However, it is important to note that the given experimental setup allows inclusion of sequential information which is not available in frequency counts alone. Further, various significance tests implicate that the extent of uniqueness and consistency is either sampling rate and/or data type dependent. Therefore, researchers should carefully consider what data types to represent using frequency models and also should consider the importance of the order in which actions are taken.

Biometric systems are known to be negatively affected by noisy samples, but little research has been done to identify and/or quantify noise in mobile device usage data. This work suggests that association rules lend some notion of “noise” in application, Bluetooth, and Wi-Fi traffic. Noise appears to correlate with items appearing with 5% or less probability. Using this threshold, the signal of the data is generally maintained across the raw data and association rules.

While there are different sampling rates seen in the research literature, the presented results suggests that intra-class similarity is maintained for all data types except for application traffic across all sampling rates. This implies that for 30, 15, and 1-minute windows, the similarity between consecutive samples should be the same. However, the means and variances for application traffic across sampling rates are very similar; further experiments are required to determine if this is ideal in this case.

Overall, experimental results suggest that frequency-based features are a satisfactory representation for mobile device usage data. It appears to be possible to distinguish between genuine and impostor data, transactions are reliably obtained over time, noise can be eliminated, it can confirm the notion of “noise”, it is beneficial to use the data types after association rule based have been generated independently.

Kruskal-Wallis H (KWH), Levene Median/Mean (LM), and ANOVA One-way (AWO) significance tests are conducted to determine if the JI is independent of the data type and sampling rate and if the centroids and spread of JI distributions are significantly different. Based on a 0.05 P-value, it is determined that the JI is time independent for all data types except application traffic; this suggests that the JI will relatively be the same regardless of the sampling rate for Bluetooth, Wi-Fi, and all data combined. Likewise, the AWQ suggests that the means are only significantly different for C03A, C15A, C5A, and C1A. The KWH also suggests that the JI is data dependent, meaning the intra-class similarity depends on the data type. LM also indicates that the variance in several subsets significantly differs; this suggests that variance will be an important factor when deciding when to apply association rules to mobile device usage data.