International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Two Factor Authentication using User Behavioural Analytics

Saurav Sen[1], Shaurya Khurana[2], Himanshu Yadav[3], Prof. Sushama A. Shirke[4]

[1,2,3]Savitribai Phule Pune University
[4]Professor, Dept. of Computer Engineering, AIT, Pune, India

---

Abstract - *Every organization loves to have new security features installed in their workstations. Usually in every organizations for each and every employee they have their personal workstations assigned to them, this brings two problems one is if a workstation is not in use and a new person wants to use it he or she cant because he is not authorized to use that. Another problem lies in security. It is often seen that employees left their workstation open, this may lead to security breach as any other person can use that workstation in absence of him and do some mal practices which is harmful to that employee. For this we aim towards developing a software product that will use the behavior of the user and will authenticate the user. We propose a framework that detects and recognizes the genuine user of that workstation using machine learning . Our approach intuitively identifies relevant features associated with behavior of user such as the speed with which is types.*

1. INTRODUCTION

User Behavior Analytics (UBA) uses big data and machine learning algorithms to assess the risk, in near-real time, of system user activity within your organization. Why is this analysis necessary? Think about it: everyday, your employees are using user credentials to access the organizations systems from the company office during regular business hours. One day you are notified that an individuals credentials were used to connect to a database server and run queries that this user has never performed before. Is a database administrator running maintenance checks or has the system been compromised? User behavior analytics can help an organization determine what normal behavior should look like within their systems and when to be cautious of unusual activity. According to the recent SANS Analytics and Intelligence Survey, only about one-third of organizations today collect user behavior monitoring data, but approximately three- fourths of respondents say they intend to start collecting this data in the future. Understandably souser behavior analytics offer visibility into potential insider threats, show early red flags for when accounts have been compromised by external attackers and are most useful to measure changes in user behavior. Ultimately, the foundation of a behavior analytics program is to understand what normal behavior looks like to catch irregularity in the system. Below are 3 key areas to focus on when establishing behavior analytics and measuring user behaviors.

• Determining human and machine behaviour Normal behaviour for accounts used by humans will look different than that of service accounts that are used to carry out automated application activity. These machine accounts usually have a large amount of permissions; however, their activity is much more predictable than human user accounts. In addition, the volume activity of automated accounts is usually much higher than human accounts. When tracking user behaviour, it is important to which type of account is being looked at when determining what unusual behaviour is.

• Track mobile device location data Mobile devices provide a great opportunity for tapping into the power of user behavior analytics. Forward-looking security programs are able to use the location tracker on smartphones as a data point in user behavior analytics. Through tracking mobile devices, security teams are able to flag any situation where an authentication is coming from a different physical location than the location of the smartphone.

• Keep tabs on machine admin accounts Companies must keep track of local machine administrator accounts in addition to active directory accounts. Cyber criminals tend to leverage these local accounts to move work their way into a system until they can break into a more critical user account. These hackers are usually successful within companies that use a standard image for rapid desktop deployment and keep local domain administrator passwords identical to simplify helpdesk requests.

User behaviour analytics are helping to transform security and fraud management by enabling organizations to detect when legitimate user accounts have been compromised by external attackers or are being abused by insiders for malicious purposes. Traditional Model vs New Approach Security Information and Event Management or SIEM, is

the traditional model that uses complex set of tools and technologies that gives a comprehensive view of the security of your IT system. It makes use of data and event information, allowing you to see patterns and trends that are normal, and alert you when there are anomalous trends and events.UEBA works the same way, only that ituses user (and entity) behavior informationto come up with what's normal and what's not. SIEM, however, is rules-based, and advanced hackers can easily work around or evade these rules. What's more, SIEM rules are designed to immediately detect threats happening in real time,

while advanced attacks are usually carried out over a span of months or years. UEBA, on the other hand, does not rely on rules. Instead, it uses risk scoring techniques and advanced algorithms, allowing it to detect anomalies over time.

One of the best practices for IT security is to use bothSIEM and UEBA to have better security and detection capabilities.

2. RELATED WORK

Comparing the documents referred,different approaches to authenticate the user using different models are analyzed. AUC. AUC is the area under the ROC which represents the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive. We also store accuracy which is true positive and true negative divided by all results.The performances of 8 different data mining algorithms - logistic regression, boosted trees, random forest, support vector machines , etc - are compared . The short analysis shows the predictive capability of machine learning algorithms for heart diseases. Possible improvements can be obtained with improved data pre-processing (outliers, variances), choice of models, parameter selection, model tuning and so on.

Shepherd, S. J[1] describes a simple, software based keyboard monitoring system for the IBM PCfor the continuous analysis of the typing characteristics of the user for the purpose of continuous authentication.By exploiting the electrical characteristics of the PG keyboard interface together with modifications to the internal system timer, very accurate measurements can be made of keystroke interval and duration, including measurements OF rollover. Rollover patterns, particularly when typing common diphthongs, can be highly characteristic of individual users and provide quite an accurate indication of the users identity. There are a number of different aspects of keystroke characteristics that can be used as identification criteria :-

• Intervals between keystrokes:-These can be analyzed on the basis of a simple mean time interval across all keystrokes or between particular pairs of keystrokes of significance such as common pairs of characters (digraphs).
• Duration of keystrokes:-Frequency of errors, This could be monitored by detecting the specific use of the delete and backspace keys.
• Force of keystrokes:-While this might give valuable additional information, no computer keyboard offers the ability to measure this quantity.
• Rate of typing:-The average number of words or characters per minute.
• Statistics of text:-The individual language use or style of a user might be analyzed but this would require significant natural language processing and would only be applicable in those situations where a reasonably large amount of text processing was being carried out. The keyboard hardware interrupt occurs once for each key depression and for each key release. The hardware scan codes associated with each key are transmitted as well so that each physical key can be identified. It is important to note that much of this information is lost after processing by the BIOS. For example, capital letters can be obtained by pressing either ship key - there is no apparent difference between them. Likewise, most keyboards have two CRTL keys and two ALT keys to suit the convenience of both right and left handed user. The BIOS keyboard processor generates exactly the same output for both shift keys or CRTL .The keyboard interrupt handler is called each time a key is pressed or released. The key associated with each operation is noted and the number of timer ticks between events is recorded. From this data, the: duration of each keystroke, the interval between, keystrokes and the overlap between keystrokes is computed. A running update of the mean and variance: of these quantities is kept and is available for display in the demonstration system.

Panasiuk, P. and Saeed K[2] modify their previous kNN algorithm and present a modification to improve the algorithm by considering key inner and interclass distinguishability. The suggested approach is tested on a large group of individ-uals with data gathered over Internet using browser-based WWW application. The obtained results are promising and encouraging for further development in this area. Data have been gathered in non-supervised way with the web-based platform. Samples consist of ve phrases that everyone has their unique features. In both language versions are the same dependencies in sample selection. Each phrase in a sample is stored in the database as a series of key events written as a text. At the beginning we read the SQL le and load it into the testing subprogram. Each keyboard event in the database is recalculated from the time a key is pressed or released to the ight time and dwell time. Flight time is the time between releasing a key and pressing the second. Dwell time is the time when a specic key is in pressed state. After loading the database le into the testing platform backspace and delete keys are removed from samples with aected keys information. Later are removed samples with dierent events count than the most common. After this, the users with less number of samples than the training set size plus at least one test sample per user are removed from the database. The next step is splitting the remained database into test and training sets. Previous steps provided constant count of training samples per user. In the following step every test sample in the remained database is being classied.

Juola, P., Noecker, J. I., Stolerman[3]developed a large corpus of keyboard behavior based on temporary workersem- ployed in a simulated office environment.

Banerjee, S. P. and Woodard[4] A behavioral biometric such as keystroke dynamics which makes use of the typing cadence of an individual can be used to strengthen exist- ingsecurity techniques effectively and cheaply.

3. PROPOSED SYSTEM

The system will help to authenticate the user using machine learning hybrid approach. The User first goes from the process of authentication by recognizing its typing pattern behavior. On successful identification the system will authenticate the user. If the system has low confidence value while recognizing the user then it will go for second authentication that is to use OTP.

4. PROBLEM DEFINITION

Problem Definition is "To build a fault tolerant, attack resistant software system using behavior analytics that will use behavior of user while using his or her workstation such as typing pattern , speed and use that to authenticate the user.

5. SYSTEM DESIGN

A. SYSTEM ARCHITECTURE

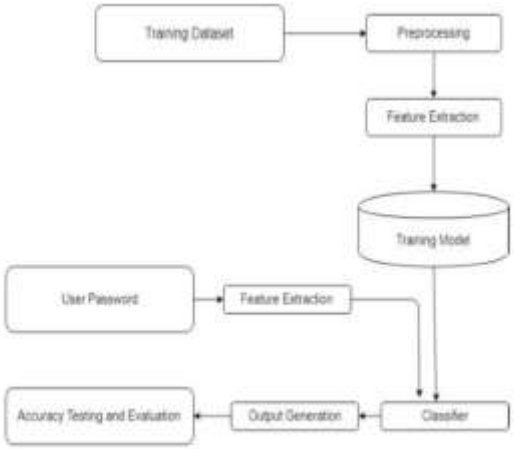

Fig 1 shows the System architecture of the proposed System.

B. SYSTEM MODULES

1) CLIENT:
a) Input Features: Client first registers and then types sentence.
b) Predict Output:
2) SERVER:
a) Classifiers:
b) Prediction:
3) ADMIN:
a) Load Data Set:
b) Classification:
c) Export to Server:

C. PROPOSED ALGORITHM

1) Start
2) If new user register for the typing pattern from admin else go to next step
3) Check for user credentials
4) If not correct reenter the user credentials else go to next step

5) Generate a random sentence
6) User types the given sentence
7) Typing pattern is identified
8) If typing pattern matches Access is granted
9) End

we present a mathematical model for kNN algorithma and show that kNN only makes use of local prior probabilities for classification.For a given query instance x t , kNN algorithm
works as follows:

$$y_t = \underset{c\in\{c_1,c_2,\ldots,c_m\}}{\arg\max} \sum_{x_i\in N(x_t,k)} E(y_i,c)$$

Where y_t is the predicted class for the query instance x t and c is the number of classes present in the data. Also

$$E(a,b) = \begin{cases} 1 & if a = b \\ 0 & else \end{cases}$$

$$N(x,k) = Set\ of\ k\ nearest\ neighbor\ of\ x$$

$$y_t = \arg\max \left\{ \sum_{x_i\in N(x_t,k)} E(y_i,c_1), \sum_{x_i\in N(x_t,k)} E(y_i,c_2), \quad \ldots \quad , \sum_{x_i\in N(x_t,k)} E(y_i,c_m) \right\}$$

$$y_t = \arg\max \left\{ \sum_{x_i\in N(x_t,k)} \frac{E(y_i,c_1)}{k}, \sum_{x_i\in N(x_t,k)} \frac{E(y_i,c_2)}{k}, \quad \ldots \quad , \sum_{x_i\in N(x_t,k)} \frac{E(y_i,c_m)}{k} \right\}$$

and we know that

$$p(c_j)_{(x_t,k)} = \sum_{x_i\in N(x_t,k)} \frac{E(y_i,c_j)}{k}$$

Where $p(c_j)_{(xt,k)}$ is the probability of occurrence of j^{th} class in the neighborhood of x t . Hence Eq.turns out to be :

$$y_t = \arg\max\{p(c_1)_{(x_t,k)}, p(c_2)_{(x_t,k)}, \ldots, p(c_m)_{(x_t,k)}\}$$

kNN algorithm uses only prior probabilities to calculate the class of the query instance. It ignores the class distribution around the neighborhood of query point.

6. INPUT DATASET

Sno	Name	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5	Rate	Duration Average
1	Shaurya	130	135	120	128	132	129	30
2	Saurav	200	210	198	180	203	211	10
3	Himanshu	250	251	240	233	255	248	50
4	Prem	168	170	172	175	165	169	40

The typing behaviour Dataset contains the unique features that can useful for identifying the user . These are
1. Time intervals between the keystrokes.
2. Rate of Typing.
3. Duration of press.
4. Error Rate

The time intervals between the keystrokes are measured as a difference of two simultaneous key presses. The rate of typing is the average number of words typed in a minute. The duration indicates the time for which the user presses the key.

The error rate is the number of time the user makes a mistake during his typing work in a time interval. These features are unique characteristics of the user which can be used to identify the user.

7. IDENTIFICATION ALGORITHMS

The machine learning algorithms can be used to identify the user uniquely to a great accuracy.The algorithms that will be used are Classification algorithms since this being a classification problem. These will work on the basis of the features taken as an input in the Data-set. Some of the algorithms that can be used are:
1. SVM (Support vector Machines)
2. KNN (K-nearest-neighbours)
3. Random Forest
4. Naive Bayes

8. ANALYSIS OF ALGORITHMS

We can infer that algorithms like nave bayes and decision tree performs better than other machine learning algorithm. It is found that instance based learning algorithm such as k-nearest neighbour performs poorly as compared to nave bayes and decision tree but performs better than rule based algorithms such as random tree. From this we can conclude that instance based learning algorithm such as k-nearest neighbour algorithm as well as nave bayes algorithm can be used for typing pattern authentication. We should avoid rule based learning algorithms such as random tree.

Learning Methods	Training	Test Accuracy
C4.5 Decision Tree	95.6%	93.3%
Naive Bayesian	93.3%	90.8%
K-star	100%	85.6%
Decision table	95.6%	81.1%
Random Tree	100%	77.8%
OneR	91.3%	75.2%
IB KNN		
$(k = 8)$	90.2%	87.4%
$(k = 7)$	91.1%	89.4%
$(k = 5)$	93.3%	91.1%
$(k = 1)$	100%	81.5%

9. CONCLUSION

User Behavior analytics is the current boom in the field of research and IT. The Proposed system is an appropriate replacement for rule based authentication system. It is expected that our approach will achieve much better results. The proposed system provides all the necessary features of a security system. This system is a cost-effective as no external hardware is used such as biometrics for a reliable authentication system. The proposed system is cost-efficient as compared to other security architectures. The architecture also benefits the users in terms of usability, and trust. We might get high performance of our classier but there is denitely scope for improvement. We evaluated our models using absolute probability thresholds, which may not be the most reliable for models where probability scoring is not well calibrated.

REFERENCES

[1] Shepherd, S. J. Continuous authentication by analysis of keyboard typing characteristics.Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system."
[2] Panasiuk, P., & Saeed, K.. A modified algorithm for user identification by his typing on the keyboard. In Image Processing and Communications Challenges 2 (pp. 113-120). Springer, Berlin, Heidelberg.
[3] Juola, P., Noecker, J. I., Stolerman, A., Ryan, M. V., Brennan, P., & Greenstadt, R. . Keyboard-behavior-based authentication. IT Professional, 15(4), 8-11
[4] Banerjee, S. P., & Woodard, D. L. Biometric authentication and identification using keystroke dynamics: A survey. Journal of Pattern Recognition Research, 7(1), 116-139.