KEYSTROKE DYNAMICS NOT WHAT YOU TYPE BUT HOW YOU TYPE

Supriya Shelke[1], Vidya Shinde[2], Sneha Pujar[3]

[1]*Student, Department Of Computer Enginnering, Pimpri Chinchwad College of Engineering, Pune*
[2]*Student, Department Of Computer Enginnering, Pimpri Chinchwad College of Engineering, Pune*
[3]*Student, Department Of Computer Enginnering, Pimpri Chinchwad College of Engineering, Pune*

ABSTRACT

Nowadays data exposure is a source of growing concern due to the widespread use of internet. Most commonly used login and password credentials may not provide enough security, as they may be easily stolen or guessed in some cases. Global access to information and resources is becoming an integral part of nearly every aspect of our lives. All the data are stored across the globe and maintained by the cloud service providers as cloud computing is growing vastly due to its advantages. Unfortunately along with this comes increased chances of malicious attacks and intrusion. Security of the data is a major concern in terms of both outsiders and insiders. The use of biometrics is a prominent alternative for user authentication, such as by the use of keystroke dynamics. This biometric technology allows the recognition of users by their typing rhythm, which can be performed using data provided by a common keyboard. Keystroke dynamics is robust against forgery attacks as well as malware detection. However, recent work has shown that typing rhythm changes over time. The proposed approach uses a host based user profiling techniques where a keystroke dynamics used for analyzing the user behavior. Along with it a retraining approach is also proposed as imposter patterns are absent at the time of registration. This retraining phase enhances security and overall performance of the system.

Keyword :- *Keystroke dynamics, cloud, authentication, security, insider threat, forgery, malware detection*

1. INTRODUCTION

The increasing use of automated information systems together with our pervasive use of computers has greatly simplified our lives, while making us overwhelmingly dependent on computers and digital networks. Technological achievements over the past decade have resulted in improved network services, particularly in the areas of performance, reliability, and availability, and have significantly reduced operating costs due to the more efficient utilization of these advancements. However, the overwhelming interest in global accessibility brought about by these advances in technology have unveiled new threats to computer system security.

Cloud computing can be generally stated as anything that involves delivering hosted services over the internet which is managed by the cloud service providers. The increased use of the cloud raises privacy concerns.

Security of the data thus becomes the major issue. A malicious insider can cause more damage to the cloud provider as well as to the users by stealing the sensitive information. For example, a cloud administrator can access virtual machine of the users and steal the information of the users without their intervention. Identity Management System (IDM) and Intrusion Detection System (IDS) are useful tools in this case. Several services are currently hosted in the Internet. Login and passwords are the most common alternatives for user access to

these services. However, it raises some questions on whether commonly used login and password credentials provides enough security, since they can be stolen or guessed. This paper investigates the use of keystroke dynamics biometric technology as an alternative. This technology recognizes users by their typing rhythm.

Keystroke dynamics is a science of studying about keystrokes that differentiate each user based on their typing speed, latency between keystrokes, and pressure applied on keys etc. Key stroke dynamics fall under non-static biometrics which will vary with time[5]. Non-static biometrics depends on several environmental, physical and biological factors. In contrast IRIS, finger prints, palm print etc comes under static biometrics which stays constant for longer duration but they require extra hardware to achieve it, which is not possible in a cloud based environment. To deal with the non-static biometric nature, different features are to be evaluated to attain proper results.

The proposed work uses a Support Vector Machine (SVM) which is one of the best known classifications and regression algorithm to date. Support Vectors (SV) that fall under different regions are separated using hyper planes linear as well as non-linear, and are achieved by adjusting different kernel functions. Researchers have proved that SVM will converge to the best possible solution in very less time. In this work a variant of the SVM called online SVM is used where the results are processed on the fly.

Although the biometric technology continues to improve, an intrinsic characteristic of this technology is that a system's error rate, for example, the false accept rate (FAR), false reject rate (FRR)

FAR is defined as the ratio of the number of times imposter gets accepted to the total number of attempts.

FRR is defined as the ratio of the number of times true user gets rejected to the total number of attempts.

2. RELEATED WORK

This paper there is a study of keystroke dynamics in a data stream context, investigating the problem of adapting user models to typing rhythm changes over time. Some classification algorithms used in the literature were evaluated and modifications to their adaptation mechanisms were proposed. In particular, the proposed modifications showed superior predictive performance.[1]

A remote authentication framework called TUBA is designed and implemented for monitoring a user's keystroke-dynamics patterns and identifying intruders. The robustness of TUBA is evaluated through comprehensive experimental evaluation including two series of simulated bots. The TUBA model can be adopted to be used for continuous and non-intrusive authentication in both, the standalone and client-server, architectures by monitoring frequently typed strings, such as usernames, passwords, email addresses, URLs, etc.[2]

Meta-analysis has been used to gauge the relative impact of single and joint attributes on the generalization behavior (performance gain) of adaptive biometric systems.[6]

By incorporating linguistic context into a keystroke-based user authentication system, we are able to improve performance, as measured by EER. Taking advantage of patterns in keystroke dynamics, we show that typists possess unique behavior which can be used to help identify the typist. We found that keystroke-based verification can be optimized using surrounding context, both of neighboring keystrokes and sub-surface linguistic context. Specifically, our experiments indicate that users type consistently based on syntactic context, and that this is more reliable than the specific word.[3].

The paper shows comprehensive survey of research efforts of the last few decades on keystroke dynamics biometrics.[4]

3. EXISTING SYSTEM

A remote authentication framework called TUBA for monitoring user's typing patterns. TUBA (Telling Human and Bot Apart) is particularly suitable for detecting extrusion in enterprises and organizations, protecting the integrity and security of hosts in collaborative environments, and as an authentication method.

Our work also suggests that certain human behaviors, namely user inputs, can be leveraged for malware detection. We give concrete examples detailing how to prevent malware forgery in such human-behavior driven security systems.

The TUBA model can be adopted to be used for the standalone and client-server, architectures by monitoring frequently typed strings, such as usernames , passwords, email addresses, URLs, etc. A database of these strings and corresponding SVM models is created during an initial training phase. After the training phase we assume TUBA to be running in the background (non-intrusively) checking the stream of typed characters for matching

strings in the database and only extracting features for evaluation against the trained models when a match occurs. When a match occurs the features of the typed string are classified as either owner or unknown.

After a number of instances are classified as unknown the user is notified of the suspicious behavior and (depending on the chosen configuration) the computer may be automatically locked, under the assumption that it's under attack. Conversely, if the majority of the instances are classified as owner then no suspicion arises.

4. PROPOSED SYSTEM

The proposed Key Stroke analyzer uses a host based user profiling technique where the monitoring subsystem is a key logger installed in each and every virtual machines and the information is gathered from both the hosts and virtual machines at regular intervals. The abnormality in the behavior will raise alarm and locks the current session.

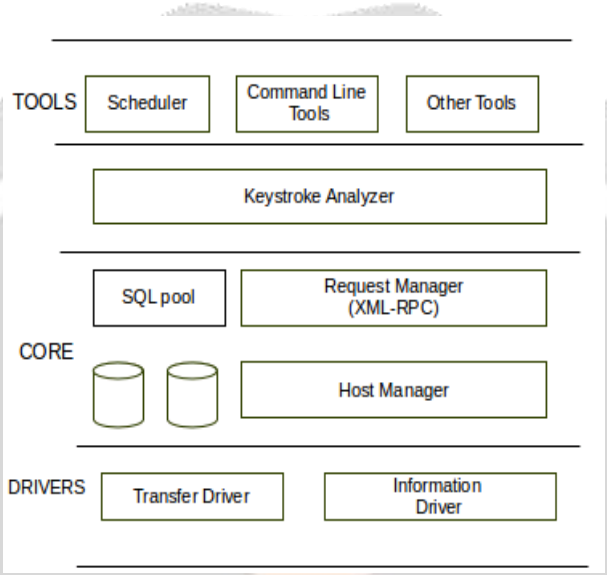

Fig. 1 : Keystroke analyzer and architecture

The proposed approach contains three phases: 1) Registration phase, 2) Validation phase and 3) Retraining phase.

4.1 Registration phase

In the registration phase, user's behavior is monitored for 3 or 4 days and data are collected in the form of key strokes along with timestamps using the previously installed key logger, assuming that in these 3 or 4 days no one else used his/her system and all the keystrokes are user specific. Data collected from the user are stored in the data store of the cloud in a timely basis. The raw data that are collected are to be processed and are arranged as a set of features before providing them as SVs to the SVM as described in processing raw data. SVM will generate a model file which will act as a user template to validate the user. That means registration phase involves User behavior, key logger and also there is a threshold calculation.

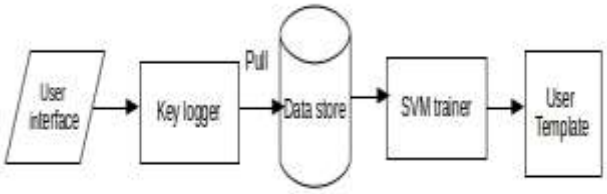

Fig. 2 : Registration phase

1) Processing raw data

Raw data collected from the user contain keys and their time stamps only. This data need to be processed before submitting it to the training, verification or retraining phases, and are to be organized into different features.

Keystroke activity generates hardware interrupts that can be time stamped and measured up to microseconds (μs) precision and even less. A keystroke activity usually consists of two actions - key press and key release which can be collected along with their time stamps. The following features can be extracted using the key press and key release values.

Dwell Time (DT): Dwell time refers to the amount of time between press and release of a single key, also called duration time and hold time and can be calculated

as :

DT =R -P , where R is the time stamp of key release and P is the time stamp of key press

Flight Time (FT) or Interval: Flight time refers to the amount of time elapsed between pressing and releasing of two successive keys, also called latency time or inters key time or interval time.

FT PP = P -P , (Press Press)

FT PR = P -R , (Press Release)

FT RP = R -P , (Release Press)

FT RR = R -R , (Release Release)

4.2 Validation phase

In the Validation phase the user is monitored and data collected is pulled to the cloud through SSH, the raw data is processed as described in the processing raw data section and compare with previously generated user template using SVM prediction method. If the current vector matches with the user template then the user is accepted by the system as true user and the data is stored in a separate data store by setting the access field to 1 for retraining. Otherwise the vectors can be either distorted or imposter patterns. Rejecting the user immediately will raise FRR which will degrade the overall system performance. A trust factor is defined to manage FAR and FRR properly through a reward penalty function.

In validation phase there is calculation of trust factor and it consist of Hy**p**erbolic SVM for classification purpose.

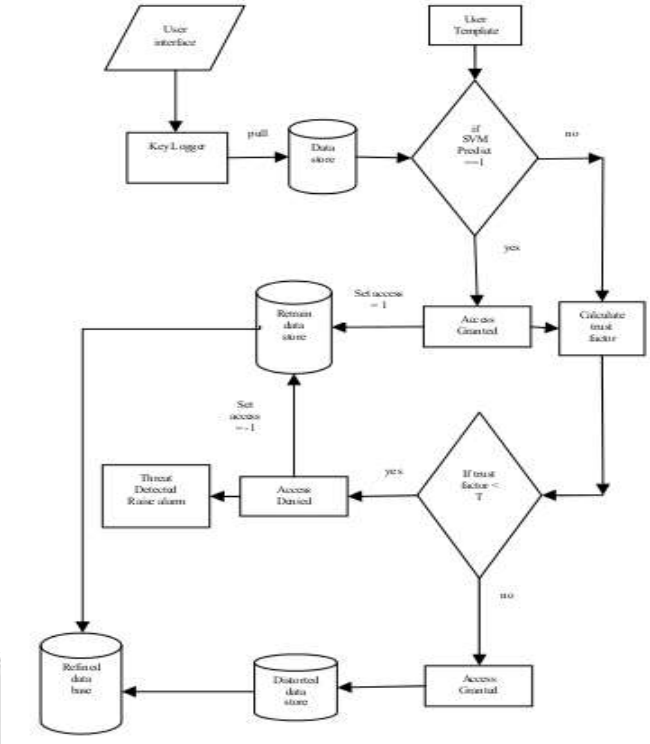

Fig. 3 : Validation phase

Because of various environmental, physical and mental conditions changes in user behavior may occur. In such a situation trust factor comes into picture. Trust factor (TF) it is one of the important factor in the continuous authentication system as the user cannot be rejected on a single failure which will increase the False Rejection Ratio (FRR). Trust factor is a reward penalty function that will increase if there is a match and decreases if there is mismatch. This reward penalty function should be chooses wisely such that FAR and FRR should be minimum.

In this approach we are using hyperbolic SVM. Reward penalty function has two types 1) Fixed reward penalty 2) Variable reward penalty. In hyperbolic SVM we are taking two planes- 1) XY Plane and 2) YZ Plane. Suppose different points are located along XY and YZ plane. There are also some points that can not be recognized whether they are on XY plane or YZ plane.

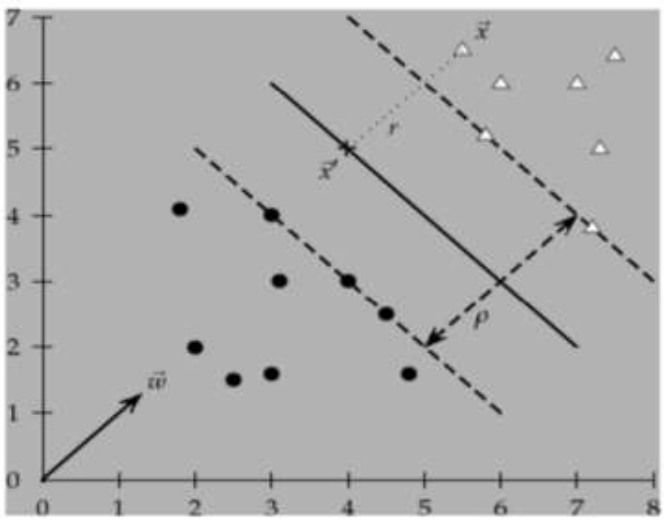

Fig.4 : Hyper planes separating SVM

The penalty method followed by keystroke analyzer is nothing but logarithmic barrier function which is represented by using following formula,

$$g(T, D) = \begin{cases} -\log(T - D) & \text{When } D < T \\ \infty & \text{Otherwise} \end{cases} \quad (1)$$

where, T is Threshold and D is Distance.

By using logarithmic barrier we calculate function f(D).Function f(D) is get minimized by using logarithmic barrier function. Variable D must be lower than some global or local threshold (T).Between the SV's and hyper plane Distance D is calculated by using formula:

$$\text{Distance} = D\left(\frac{\vec{W}}{|W|}\right)$$

From above fig.4 we are taking one point which is having minimum distance from hyperplane. Shortest distance D of that point is perpendicular to the plane and hence parallel to a hyper plane ($\vec{W}$).Vector can be represented as ($\vec{W}$) and determined as :

Vector=Magnitude + Direction

The unit vector in this direction is calculated

as $\frac{\vec{W}}{|W|}$. Nearest point from hyperplane on Z axis from SVM locus can be represented as $\vec{X}'$ and it can be determine by using formula:

$$\vec{X}' = \vec{X} - yD\left(\frac{\vec{W}}{|W|}\right) \quad (2)$$

Sign of "y" is changes for two cases of $\vec{X}$ being on either side of the decision surface and should be always need to be positive 1(+1) because in this approach needs to calculate the distance from the positive hyper plane, then from (1)

$$\vec{X}' = \vec{X} - D\left(\frac{\vec{W}}{|W|}\right) \quad (3)$$

Since $\vec{X}'$ lies on the decision boundary, it satisfies the equation

$$\vec{W}^T\vec{X}' + b = 0 \quad (4)$$

From (3) and (4),

$$\vec{W}^T\left(\vec{X} - D\left(\frac{\vec{W}}{|W|}\right)\right) + b = 0 \quad (5)$$

Now we can calculate TF as follows (1) as:

$$TF=\begin{cases} TF - \log(T - D) \text{ if match! } = \text{true user(penalty)} \\ TF + \log\left(\frac{2}{|w|}\right) \text{match } == \text{true user(reward)} \end{cases}$$

If TF is less than Threshold then access denied otherwise access granted.

4.2 Retraining phase

The retraining phase is same as of validation phase but the efficiency of the retraining lies in its capacity to retrain the new data by modifying the existing model file. An online SVM algorithm can obtain this by altering the direction searches and will converge to the known SVM solution. The advantages by using this method in the keystroke analyzer are

• The training timing required is optimized as Support Vectors (SV) can be divided into sub vectors and trained through break points.

• The algorithm will update the samples only when there is a violation in Karush-Kuhn-Tucker (KKT) conditions until the new data point satisfies the optimality condition.

• Online algorithms require very less memory as the vectors are processed one by one and are discarded after examination.

Fig. 5 : Retraining Phase

5. CONCLUSION

As nowadays use of internet is widely increased data exposure became a source of growing concern. Keystroke dynamics facilitates a natural and cost effective way for security and access protection of computers and mobile devices. It also allows for continuous authentication by monitoring a user's typing behavior during the entire login session without any interruption to the user's routine work. The use of keyboards for personal identification had been studied even before personal computers were introduced. It has been attracting increasing attention and interests as our increasing dependency on computers and mobile devices to store private and sensitive information demands strong security protection.

Keystroke dynamics has unmatched usability tremendous potential for cyber security applications. In cloud the insider threat has continued to be the biggest problem to date. The proposed work shows better results in mitigating the insider threat in the presence of a masquerader and as well as provides authentication to the user. In addition, the proposed approach does not require any extra hardware and there is no need of any modification in the existing cloud infrastructure for implementation. In future, the efficiency of the proposed approach will be increased by combining it with the other behavioral techniques like search and command sequence analysis.

6. REFERENCES

[1] IEEE, Keystroke-Dynamics Authentication Against Synthetic Forgeries, Deian Stefan, Danfeng(Daphne) Yao

[2] P. H. Pisani, A. C. Lorena, and A. C. de Carvalho, "Adaptive approach for keystroke dynamics," Journal of Intelligent & Robotic Systems, pp. 1-17,2014

[3] Adam Goodkind, David Guy Brizan, Andrew Rosenberg "Improvements to Keystroke-Based Authentication By Adding Linguistic Context"

[4] *Teh, Pin Shen, Andrew Beng Jin Teoh, and Shigang Yue. "A survey of keystroke dynamics biometrics." The Scientific World Journal 2013*
[5] IEEE Keystroke dynamics as a biometric for authentication, Fabian Monrose, Aviel D. Rubin, March 1999.
[6] N. Poh, A. Rattani, and F. Roli, "Critical analysis of adaptive biometric systems," Biometrics, TET, vo!. I, no. 4, pp. 179-187, 2012.
[7] wikipedia[Online] Avaliable http://*en.wikipedia.org/wiki/Keystroke-dynamics*
[8] *Zeljka Zorz [Online] Avaliable: http://www.net-security.org/article.php?id=1610*