JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

K-Nearest Neighbour classification for classifying user with high/low risk for implementing risk-based authentication.

[1]Vanita Gundu Sutar, [2]Asst. Prof. Deepali Jadhav

Department of Computer Science and Engineering
KIT's College of Engineering (Autonomous), Kolhapur, India

Abstract : Risk based authentication is an approach to authenticate user based on risk score assigned to used based on previous authentication history. It is adaptive authentication in which based on risk score user is asked for additional authentication steps, if risk score is low, user will be allowed to login frictionless.

The purpose of this study is to classify users based on risk in authentication as HIGH/LOW risk using K nearest neighbour algorithm and check the accuracy results of the KNN algorithm for the risk based authentication classification of genuine vs fraud login attempt.

IndexTerms **– Risk based authentication, KNN, classification algorithms**

I. INTRODUCTION

Risk based Authentication also known as adaptive authentication solutions assign a risk score based on user login behaviour. Additional rules also can be applied to assign risk score. Machine learning algorithms can be used to learn user behaviour to build a user profile for login patterns.

Risk based authentication asks users less authentication information whose behaviour is in certain expected way (same login device/IP for most of the logins, same geo locations for most of the logins). This will result in lesser friction for user authentication without compromising on security. If user login behaviour is not similar to usual, user can be requested with more authentication information in addition to general login information (MFA, OTP etc). Machine learning classification algorithms can be used to classify whether a user attempting to login is genuine or fraud based on risk analysis.

K nearest neighbor algorithm for classification

The K-nearest neighbors (KNN) algorithm is supervised machine learning algorithms. KNN is easy to implement in its basic form, also performs complex classification tasks. It is called lazy learning algorithm which means it is not having a specialized training phase. It uses the entire data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, it doesn't assume anything about the underlying data. This is useful feature because the realworld data doesn't really follow any theoretical assumption e.g., linear-separability, uniform distribution, etc. For this type of problem, KNN will be useful. With the help of K-NN, we can easily identify the category high risk /low risk authentication of a given login attempts dataset.

II. LITERATURE REVIEW

Most of the RBA-related research is focused on evaluating the reliability and robustness of certain features.

From Reference [1], below features will give highly distinguishable information for RBA –

- IP Address
- Geolocation
- User String (subdivided into browser, operating system (OS) and version)
- Language
- Login time

There are additional features - Canvas fingerprinting and evercookies which were not considered in study by the reference paper [1].

Reference paper [2] suggests on - How many user sessions need to be captured and stored in the login history to achieve a stable and reliable RBA setup -

EXTEND (EXTEND model uses multiple features for RBA) required ten entries to block 99.92 % of attacks. The SIMPLE (single feature-IP Address) models partly did not fulfill the requirement. Based on the results, storing one entry is already sufficient for a stable setup that blocks more than 99.45% of targeted attackers with the EXTEND model. Additional feature of RTT – round trip time is used RTT< 10 ms as Attackers need access to a device physically located inside the victim's location to forge this feature.

There are other features which are reliable, but client side oriented, hence spoof able - OS name and version, Browser major version.

Feature selection:

Reference paper [1] concludes on the features that will give highly distinguishable information for Risk based authentication. Referring to it, will be using IP Address, User Agent String(device), Location, Failed Login Attempts as features for classifying HIGH/LOW risk for the login attempt instance.

Data pre-processing by rescaling: KNN performs much better if all of the data has the same scale. Normalizing your data to the range [0, 1] will give better results. The input features data to algorithm will be created by having data in range [0,1] i.e. for a new login attempt instance capture user id, username IP address etc. and query in the login attempt history table. Create input feature row like - isdevice_id_changed, islocation_changed, isip_address_changed, islast_login_sucess in terms of 0 if the values remain same with the login history and 1 if they are changed in new login instance.

Lower Dimensionality: KNN is suited for less dimensional data. So only minimal features are selected which are most significant for detecting anomaly in login behavior.

III. THEORETICAL UNDERSTANDING

The training phase of K-Nearest Neighbour classification is much faster compared to other classification algorithms. There is no need to train a model for generalization, hence KNN is useful in this case. Also, K-nearest neighbour is an instance-based algorithm.

Let's say a new instance x is provided for prediction, the entire training dataset is searched for K most matching instances (called as neighbours) summarize the predicted output for those K instances.

Euclidean distance measure is used determine which of the K instances in the training dataset are most similar to a new input.

Euclidean distance is the square root of the sum of the squared differences between a new point (p) and an existing point (pi) across all input attributes j.

$$\text{Euclidean Distance } (p, p_i) = \text{sqrt}(\text{sum}((p_j - p_{ij})^2))$$

IV. PROPOSED WORK

KNN algorithm to be evaluated in terms of accuracy using metrics Precision, Recall and F1-score. The results can be compared with other algorithms like SVM, Random Forest algorithms to choose the best fit algorithm for the given problem of classification for RBA.

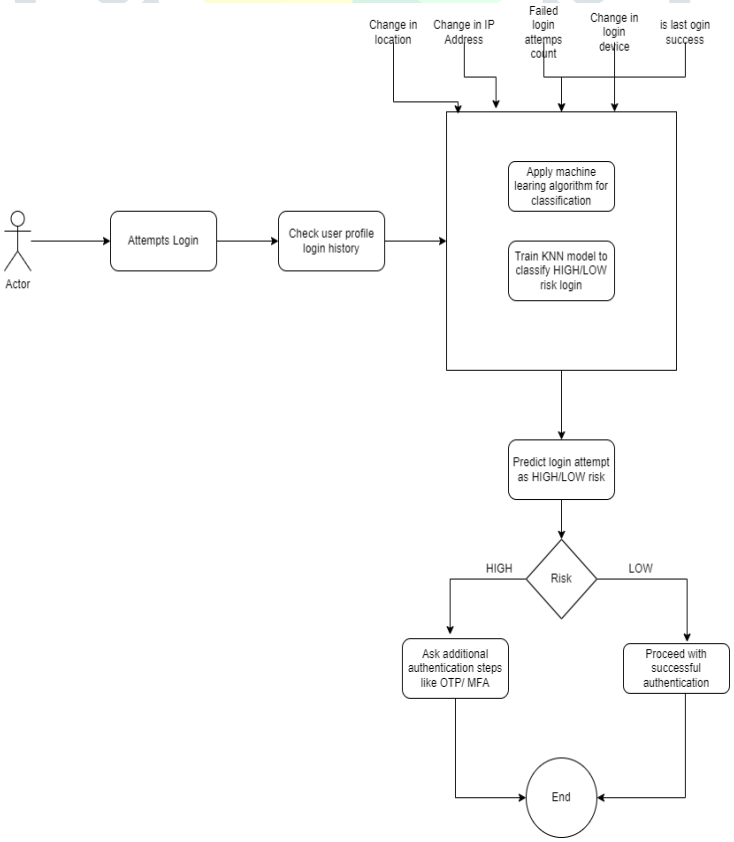

Fig.a Flow diagram for risk level classification

VI REFERENCES

[1] Is This Really You? An Empirical Study on Risk-Based Authentication Applied in the Wild Stephan Wiefling1(B) , Luigi Lo Iacono1 , and Markus D¨urmuth2 1 TH K¨oln - University of Applied Sciences, Cologne, Germany {stephan.wiefling,luigi.lo iacono}@th-koeln.de 2 Ruhr University Bochum, Bochum, Germany markus.duermuth@rub.d.

[2] A Systematic Approach for a Secure Authentication System H A Gautham, Ramakanth Kumar

[3] What's in Score for Website Users: A Data-Driven Long-Term Study on Risk-Based Authentication Characteristics Stephan Wiefling1,2 , Markus Dürmuth2 , and Luigi Lo Iacono1 1 H-BRS University of Applied Sciences, Sankt Augustin Germany {stephan.wiefling,luigi.lo iacono}@h-brs.de 2 Ruhr University Bochum, Bochum, Germany {stephan.wiefling,markus.duermuth}@rub.de

[4]https://www.researchgate.net/figure/Setup-of-study-2-showing-the-probed-features-We-tested-all-possible-combinations-ie_tbl3_333557618

[5] https://www.ee.co.za/article/risk-based-authentication-convenience-security.html

[6] https://riskbasedauthentication.org/state-of-practice/results/rba-models-algorithms/

[7] https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn