Role of Machine Learning in change of Identity and Access Management: A proposed design of IAM for University System

Subhrodip Roy Chowdhury
College of Engineering and Technology
East Carolina University
Greenville, NC 27858
Email: roychowdhurys18@students.ecu.edu

***Abstract*—In recent years, machine learning is revolutionizing almost every field of science due to its unique properties like adaptability, scalability and ability to handle unknown challenges. It is targeted towards reducing human effort and intervention. Identity and Access Management (IAM) is also not an exception and currently it is in a critical juncture to handle social media, online transactions, cloud and web technologies and IoT devices. Reviewing the literature for the current Identity and Access Management systems and utilizing the knowledge appropriately to extend the existing models in the light of Artificial Intelligence especially Machine Learning will be an immensely valuable addition. Access Control system is currently moving from Role Based Access Control (RBAC) models to Attribute Based Access Control (ABAC) models. In ABAC models, environment attribute can be easily modeled with machine learning to provide user access. Machine learning can also be used to determine the role and policies by combining both RBAC and ABAC models. The present strategies like credential based, two-factor or token based authentication is not sufficient enough for user access. Malicious activities should be tracked during and after authentication. Hence, there is a research gap which can be bridged by effectively using machine learning to study user behavior and work pattern and therefore monitor activities even after user access is granted. This paper proposes a design of Identity and Access Management for University System using machine learning for access control. It also highlights the advantages and disadvantages of using machine learning techniques in ABAC models and suggest a discretionary mechanism before the system is matured and stabilized by itself.**

Keywords—IAM, RBAC, ABAC, IDS, XACML, NSA, UBA, SIEM

I. INTRODUCTION

IAM has been deeply influenced by the development of access management models such as role-based access control (RBAC),[1] decentralized trust management (DTM),[2] and attribute-based access control (ABAC) [3]. These and similar models have improved management efficiency and enabled new levels of automation.

In an important early study of intrusion [4], Anderson postulated that one could, with reasonable confidence, distinguish between an outside attacker and a legitimate user. Patterns of legitimate user behavior can be established by observing past history, and significant deviation from such patterns can be detected. Anderson suggests that the task of detecting an inside attacker (a legitimate user acting in an unauthorized fashion) is more difficult, in that the distinction between abnormal and normal behavior may be small. Anderson concluded that such violations would be undetectable solely through the search for anomalous behavior. However, insider behavior might nevertheless be detectable by intelligent definition of the class of conditions that suggest unauthorized use. These observations, which were made in 1980, remain true today.

Machine-learning approaches use data mining techniques to automatically develop a model using the labeled normal training data. This model is then able to classify subsequently observed data as either normal or anomalous.

There are multiple challenges the to Identity and Access Management system to identify the subject and provide them proper access to the requested objects. We have to understand the request or access scope and environment before providing the access. Only role based access system would fail to consider environmental facts. So this is the time to move from RBAC to ABAC which current market leader products are not using full faze because of performance issue. Also we need to make the IAM system learn by itself and decide on runtime because every day the vulnerability and threat mechanism are changing and along with external threat the internal threat and policy violation are increasing. Since Edward Snowden's release of over 1.7 million classified NSA documents, trusted identities and privileges require re-examination. If we consider inside threat then need to consider the environmental factor more seriously. We will go through the different section of Identity and Access Management system to identify the area where we can implement machine learning and how we can reduce the risk.

An identity manager can also ensure users are automatically assigned roles containing appropriate entitlements as part of onboarding. It provides an auditable record and enforces access management policies across an enterprise. An effective user provisioning/deprovisioning solution also automatically removes access privileges upon termination. It guarantees a user's access is removed in a timely manner. Also we need to keep in mind that proper authorize user should get access to system so that work hours of organization don't get affected.

Experience-Based Access Management promises broad applicability across many domains after assessment of domain-specific risks to judge trade-offs such as the balance of false positives and negatives [5]. Provision proper user based on attributes, tasks, time and create alert for any suspicious access and deactivate provision temporarily on suspicious activities.

II. Related Work

In general, typical IAM systems are built on three pillars: processes, technologies and policies [6]. Core identity lifecycle processes like user deprovisioning or access privilege management are implemented using available automation technologies. Existing products offer a variety of functionalities like identity directories for data storage, provisioning engines for user management or workflow capabilities. Both processes and technologies are controlled by a set of company-specific policies. These policies control technological aspects like data synchronization or data storage. At the same time, they are responsible for process-related aspects like access privilege management, provisioning processes, and security management within the IAM [7]. Policy management commonly still needs to be carried out manually by IT administrators with hardly any means for structured policy definition or ongoing policy management being available. Moreover, only static data is employed (e.g. department of an employee), letting valuable data lie fallow. As a result, only a small number of basic policies are defined and implemented in practice. These policies are commonly extracted from partly documented internal regulations and requirements and remain unchanged during system operation. It is mandatory that policies evolve over time in order to reflect organizational and technological changes within a company. In the field of policy management, researchers have proposed a variety of top-down and bottom-up policy detection approaches. Besides general policy mining approaches, the research community recently focused on mining attribute policies for ABAC [8] in order to ease the migration from traditional access control models such as RBAC.

Access-control policies are used to govern the various types of access that different entities may have to information. As a result of the complexity introduced by hard coding policies into programs [9], an increasing trend is to define policies in a standardized specification language such as XACML [10] and integrate the policies with applications through the use of a Policy Decision Point (PDP). Recently several tools have been developed to verify specific properties of a given XACML policy [11]. Identify discrepancies between the policy specification and the true desires of the policy authors by finding specific requests that are likely bug exposing. These observations are used as input, in the form of request-response pairs, to a particular class of machine learning algorithms called classification learning. The output of the machine learning algorithms is essentially a summary of the policy in the form of inferred properties that may not be true for all requests but are true for most requests. Evan and Tao have integrated Sun's XACML implementation [12] and a collection of machine learning algorithms for data mining tasks into a tool that implements our approach through request generation, request evaluation, and policy property inference [13].

The focus areas would be as described by Matthias et al in Adaptive identity and access management [7]

- Minimizing efforts to define an initial set of policies.
- Improve the quality and adaptability of input parameters of policies.
- Providing tool support to enable human IAM engineers to execute policy modelling and refinement.
- Integrating both actual authorization usage data and business knowledge.
- Improving IT security through continuous refinement of policies based on actual employee behavior.

Machine learning techniques offer potential solutions that can be employed for resolving such challenging and complex situations due to their ability to adapt quickly to new and unknown circumstances.

Cloud-based access control solutions are growing incredibly quickly, with recent research forecasting a compound annual growth rate of around 30 percent worldwide. These platforms are part of the broader trend toward managed security solutions, which benefits the end user in several ways, while also providing the VAR with a compelling product offering [14].

III. Conceptual Overview

We'll consider a university IAM system as it has different types of internal and external users who access different objects or resources. We'll analyze the provisioning, access mechanism, resource access request, deactivation flow and deprovisioning mechanism. There are different types of end users with different roles and responsibilities in university system. Student, faculty, employee, guest user, parents, workshop users etc. Some of these users are tightly coupled or internal users and some are loosely coupled or external users. This is the reason the IAM of university require a robust IAM system which would be able to validate the user current role and status before providing them to access to any specific object. Distance education system enables faculty and students to join classroom from different parts of world. They also need lab machines and different system access from multiple location. On the other hand employees work on campus most of the time except some IT support staff. Once the users are provisioned to any specific object then need to verify their current status before providing access. As a part of regular evaluation need to deactivate and deprovision users who are currently not in proper role. If we consider employee or faculty then providing access

is more critical and need to consider the environmental factors. Specially tracking contractor what data they are accessing and when they are accessing. Also there are different type of software used in a university based on demand of different department. Provisioning and deprovision users in multiple different systems and maintain action mapping comparing object and subject attributes on runtime can be done more accurately and robustly by machine learning. The ABAC system has access control policies which can be used to prepare test data. The test data can be plot in multivariate linear regression graph. The reason of choosing multivariate as we need to consider object attributes, subject attributes, environmental attributes. After providing test data the machine learning would be able to evaluate users on runtime based on their properties.

User authentication is performed by Active Directory or Kerberos and then it comes for authorization. We are planning to implement Machine Learning in the authorization process so that only valid user get access to system on proper time from proper location. As part of machine learning we need to choose mechanism how the attributes and action policies can be learned through test data and evaluate user access request at run time quickly. We need to keep in mind about the processing time as it relates with the performance of system, users and work hour utilization. The training data preparation from ABAC model is easier as we can map the access rules based on attribute values. Once all the access rules are defined these will help to prepare the test data. The test data can be plot to multivariate linear regression. We are choosing linear regression as this is faster evaluation process. The reason to choose multivariate as we have to consider multiple (subject, object and environment) complex attributes.

Rule-based systems have their own disadvantages:

1. Time zone challenges – Consider that there is a rule which allows employees to log in during business hours. However, this rule does not work well when the employee travels to a different time zone.
2. Rule-based systems increase the friction in terms of user experience. Let us take an example – say you have a rule which steps up the authentication if your last login location and current location do not match. While it seems to work for most cases without issues, consider the situation encountered by your sales person who travels frequently. The sales person almost always ends up being compelled to use multi-factor authentication.
3. There is no one rule that fits all in the organization. Learning-based results in growing number of Rule which eventually becomes a management nightmare and results in security holes.
4. Handling rule conflicts – how to handle scenarios wherein Rule #1 indicates opposite of Rule #2.

A machine learning based system will correctly profile a part time faculty who is constantly traveling and hence ends up accessing the system from different geographies and at different times. His profile is different from a full time on-campus employee who ends up logging in from the same geography and typically in the same time window most of the times. If the visiting faculty has logged in from India and then in the next half an hour he is trying to log in from the USA, then the system will be smart enough to determine that it is not possible to travel that far in the given time. Thus it will consider the access as suspicious and ask for an additional factor in the authentication process

A machine learning based IAM system will enable organizations to do away with some of these rules. The system will itself learn based on past patterns and accordingly, it can decide how to grant someone system access under different conditions or require trust elevation. There are two things that make machine learning ideal for this process.

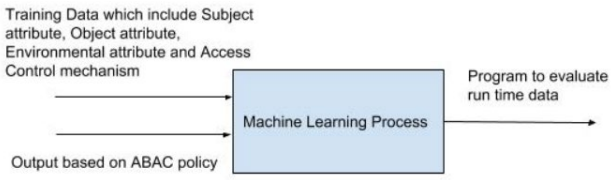

Fig. 1. Outline of Machine Learning process

User behavior analytics (UBA) as defined by Gartner is a cybersecurity process about detection of insider threats, targeted attacks, and financial fraud. UBA solutions look at patterns of human behavior, and then apply algorithms and statistical analysis to detect meaningful anomalies from those patterns—anomalies that indicate potential threats.

Security Information and Event Management (SIEM) system can be rules-based or employ a statistical correlation engine to establish relationships between event log entries. Advanced SIEMs have evolved to include user and entity behavior analytics (UEBA) and security orchestration and automated response (SOAR). In the computer security market, there are many vendors for UEBA applications.

The UBA and SIEM would add logging and monitoring system more robust. The analysis of user behavior based on prior access and work habit would predict system if user is not any specific resource long time or accessing any resource even on vacation time.

IV. EVALUATION

First we are going to identify the different subject, object and environment attributes of University system.

Subject Attributes:

1. Working hour - define the working hour of the subject. If the subject work only in office hour or any time of day.

2. Working location - defines if the subject is on-campus or off-campus user.
3. Roles - defines the roles of subject. This field considers users academic plan, sequence, employee class, department etc. which defines the user need to accomplish any work.
4. Access State - defines the current status of user (if user is active or inactive). There is also multiple values possible like disabled, deprovisioned, policy violation flagged.
5. User behavior analytics (UBA) - define user behavior data analytics. This field can be used to capture if user trying to access any resource or object out of regular behavior or nature of work. This is a derived field and need to feed constantly after users starts using the system.

Object Attributes:

1. Affiliation - define what type of employee or faculty or student affiliation needed to access the object. There are multiple affiliations options possible for students. E.g current, graduate, degree, non-degree etc. Different Employee affiliation also should classify the full time, part time, contractor etc.
2. Access Hour - define the day and time object can be accessed. If there is any execute permission then if there is any limitation on execution time. E.g. some of the production change script should be executed on defined off hours.
3. Access Location - define from which location object can be accessed. If the object access limited to on-campus.

Environment Attributes:

1. Current Date and Time - verify user access request date and time
2. Location - verify if user currently present in on-campus or outside-campus
3. System status - verify the system status flag if it is under any threat or regular mode. Also can be defined like off hour or work hour.

Action:

1. Read/write/execute/login

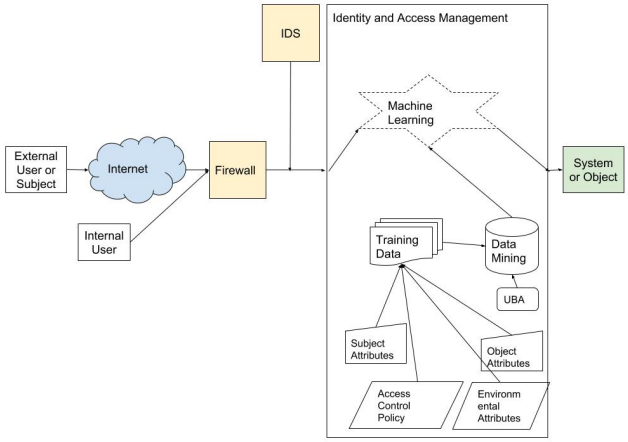

Fig. 2. Outline of IAM

Identity and Access Management has 2 parts, authentication and authorization. The authentication part is to authenticate user either by password or by tokens. We are assuming users is getting authenticated successfully and then waiting for authorization and we will analyze the different flows of authorization.

Auto provision flow would work as below:

New user - Trusted source recon create new user in IAM. Then evaluation of the user is performed to determine user role and access. The machine learning will get the user attributes for first time. But with the minimum mandatory user attributes along will subject and environment attribute would help to plot user in multivariate linear regression. The machine learning would determine users' access.

R1: can_provision(u, m, e) ←

(Roles (u) = Full Time Employee) V (Affiliation (m) = Full Time) V (System Status (e) = Regular)

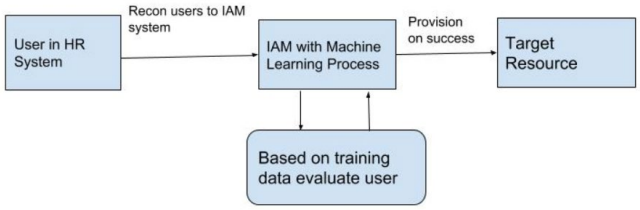

Fig. 3. Outline of Provisioning Flow

If there is any issue with Machine Learning then it is easy derive type of user based on roles which is RBAC system. This is faster process to provide user initial access. A flag can be set which would help the system to choose which flow (ABAC or RBAC) to choose. But this is for initial phase of the implementation to avoid user work hour utilization.

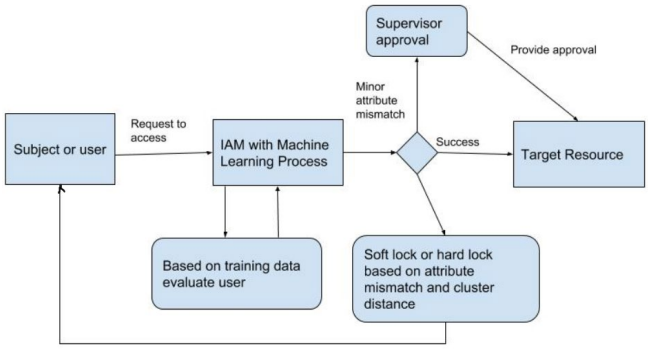

Fig. 4. Outline of Access Request Flow

Existing user role change - The machine learning would expect all subject attribute present and modified accordingly. So all the attributes can be used to change user position in the graph. The clustering method can be used here to identify the dimension of change. This can be reported to higher authority in case dimension of user role change is higher than threshold level. The resources provisioned and deprovisioned status would change with the user role change. The access control policy would play a bigger role to set the new position of user in the multivariate linear graph.

The access management system should check user attributes, subject attributes, environment attributes and access control policies based on multivariate linear regression before providing access.

The valid user while trying to access any resource but getting denied for any environment value change then the user can be soft deactivated. On supervisor discretion user can gain access. But if user gets denied for any user attribute change then it would be hard deactivation. Which can be achieved by changing policy violation flag.

If a user fails a login once on a Monday morning from their home or work, it's more likely to be a human error, and presents a relatively low risk. There's no need to flag this to the security team, but this information should be logged and accessible for later analysis. If this is happening on a Wednesday afternoon, and multiple attempts are being made from an unregistered machine that's hundreds of miles away, each factor increases the risk.

R2: can_access (u, m, e) ←

(Working hour (u) = Office Hour) V (Working location (u) = on-campus) V (Roles (u) = Full Time Employee) V (Access State (u) = Active) V (Affiliation (m) = Full Time) V (Access Hour (m) = Office Hour) V (Access Location (m) = on-campus) V (System Status (e) = Regular) V (Current Date and Time (e) = Office Hour) V (Current Location (e) = on-campus)

Auto Deprovision - Deprovisioning users for any resource based on subject or object attribute change is a regular process. There should be two different job process one would check change in user attribute and other would check change in subject attribute. Based on the subject or object attribute change the set of users already provisioned should be re-evaluated through machine learning.

Get the square root of distance of user from regular point or last point of same user from graph then derive the distance. If it is small then send dual authentication, if more than send for manager approval. This will help to reduce prod data change in odd hour.

V. CONCLUSION

A key disadvantage is that this process typically requires significant time and computational resources. Once the model is generated however, subsequent analysis is generally fairly efficient. The advantages of the machine-learning approaches include their flexibility, adaptability, and ability to capture interdependencies between the observed metrics. Their disadvantages include their dependency on assumptions about accepted behavior for a system, their currently unacceptably high false alarm rate, and their high resource cost.

Despite all the facts and drawbacks the Machine Learning is providing huge benefit to IAM reducing IT Administrator effort and automated alert system. Specifically handing the insider attack. As the world moving from desktop to laptop to IoT devices and cloud based software it is not easy to maintain roles and policy with changing system and infrastructure. So only solution is to adopt Machine Learning based IAM and provide accurate training data set to reduce false positive.

REFERENCES

[1] D.F. Ferraiolo, D.R. Kuhn, and R. Chandramouli, "Role-Based Access Control", Artech House, 2003.

[2] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized Trust Management," Proc. 1996 IEEE Symp. Security and Privacy, IEEE CS Press, 1996, pp. 164–173.

[3] L. Wang, D. Wijesekera, and S. Jajodia, "A Logic-Based Framework for Attribute Based Access Control," Proc. ACM Formal Methods in Software Eng. Workshop, ACM Press, 2004, pp. 45–55.

[4] Anderson, J. "Computer Security Threat Monitoring and Surveillance." Fort - Washington, PA: James P. Anderson Co., April 1980.

[5] Carl A. Gunter., David M. Liebovitz, Bradley Malin

"Experience-Based Access Management - A Life-Cycle Framework for Identity and Access Management Systems."

[6] L Fuchs, G Pernul, in The Second International Conference on Availability, Reliability and Security, 2007: ARES 2007. Supporting compliant and secure user handling—a structured approach for in-house identity management (IEEE Computer Society, Los Alamitos, 2007), pp. 374–384

[7] Matthias Hummer , Michael Kunz, Michael Netter, Ludwig Fuchs and Günther Pernul, "Adaptive identity and access management—contextual data based policies", EURASIP Journal on Information Security (2016) 2016:19

DOI 10.1186/s13635-016-0043-2

[8] VC Hu, D Ferraiolo, R Kuhn, A Schnitzer, K Sandlin, R Miller, K Scarfone, "Guide to attribute based access control (ABAC) definition and considerations". NIST Spec. Publ. 800, 162 (2014)

[9] K. Fisler, S. Krishnamurthi, L. Meyerovich, and

M. Tschantz. “Verification of change-impact analysis of access-control policies”. In International Conference on Software Engineering, pages 196–205, 2005.

[10] OASIS. OASIS eXtensible Access Markup Language (XACML). Published Standard, 2005

[11] Sun Microsystems. Sun’s XACML Implementation. Sourceforge, 2005.

[12] I. H. Witten and E. Frank. “Data Mining: Practical Machine Learning Tools and Techniques”. Morgan Kaufmann, 2005.

[13] E. Martin, Tao Xie, “Inferring access-control policy properties via machine learning”, Seventh IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'06)

Date of Conference: 5-7 June 2006 Print ISBN: 0-7695-2598-9

[14] Ingram Micro, “Unlocking New Potential: The Future of Access Control” www.ingrammicroadvisor.com/unlocking-new-potential-the-future-of-access-control