Measuring of Data Quality in KYC Using Anomaly Detection Techniques

Tejakshi N S*, Vyshnavi M K, Manjuprasad B

Department of Computer Science and Engineering GSSSIETW, MYSURU, INDIA

DOI: https://doi.org/10.21467/proceedings.1.39

* Corresponding author email: tejakshi29@gmail.com

Abstract

Intrusion detection has gain a broad attention and become a fertile field for several researches, and still being the subject of widespread interest by researchers. The intrusion detection community still confronts difficult problems even after many years of research. Reducing the large number of false alerts during the process of detecting unknown attack patterns remains unresolved problem. However, several research results recently have shown that there are potential solutions to this problem. Anomaly detection is a key issue of intrusion detection in which perturbations of normal behavior indicates a presence of intended or unintended induced attacks, faults, defects and others. This paper presents an overview of research directions for applying supervised and unsupervised methods for managing the problem of anomaly detection. The references cited will cover the major theoretical issues, guiding the researcher in interesting research directions.

Index Terms- Anomaly Detection, Intrusion Detection, KYC

1 INTRODUCTION

Anomaly detection is important when the abnormal behavior in the dataset provides significant information about the system. Anomalies can be caused bymalicious activities, instrumentation errors, and human errors [1]. Anomaly detection is an important problem in several application domains such as credit card fraud detection in financial systems, intrusion detection in communication systems, and contagious disease detection in public health data. Intrusion detection is probably the most well-known application of anomaly detection [2] [3]. In this application scenario, network traffic and server applications are monitored. Potential intrusion attempts, and exploits should then be identified using anomaly detection algorithms. Besides this network-based intrusion detection, also host-based intrusion detection systems are available, commonly using system call data of running computers. Most security vendors often call anomaly detection in this context behavioural analysis [4]. An important challenge in these often-commercial Intrusion Detection Systems (IDS) is the huge amount of data to be processed in near real-time. For this reason, these systems typically use simple but fast anomaly detection algorithms. Intrusion detection systems are also a good example where anomaly detection complements traditional rule-based systems: They typically use pattern matching for the fast and reliable detection of known threats while an additional anomaly detection module tries to identify yet unknown suspicious activity. Anomaly detection methods:

- **Supervised Anomaly Detection** describes the setup where the data comprises of fully labeled training and test data sets. An ordinary classifier can be trained first and applied afterwards. This scenario is very similar to traditional pattern recognition with the exception that classes are typically strongly unbalanced. Not all classification algorithms suit therefore perfectly for this task. For example, decision trees like C4.5 [5] cannot deal well with unbalanced data, whereas Support Vector Machines (SVM) [6] or Artificial Neural Networks (ANN) [7] should perform better. However, this setup is practically not very relevant due to the assumption that anomalies are known and labeled correctly. For many applications, anomalies are not known in advance or may occur spontaneously as novelties during the test phase.

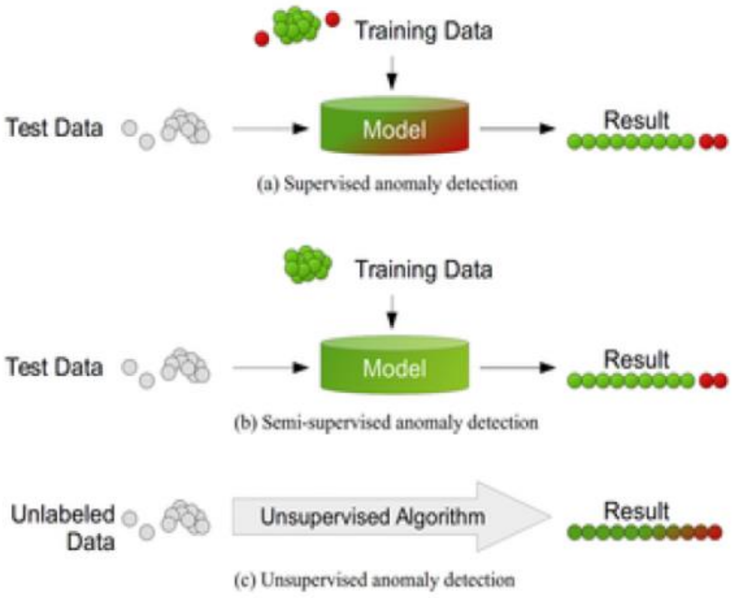


Fig 1: Anomaly Detection Methods [7]

- **Semi-supervised Anomaly Detection** also uses training and test datasets, whereas training data only consists of normal data without any anomalies. The basic idea is, that a model of the normal class is learned and anomalies can be detected afterwards by deviating from that model. This idea is also known as "one-class" classification [8]. Well-known algorithms are One-class SVMs [9] and auto encoders [10]. Of course, in general any density estimation method can be used to model the probability density function of the normal classes, such as Gaussian Mixture Models [11] or Kernel Density Estimation [12].

- **Unsupervised Anomaly Detection** is the most flexible setup which does not require any labels. Furthermore, there is also no distinction between training and a test dataset. The idea is that an unsupervised anomaly detection algorithm scores the data solely based on intrinsic properties of the dataset. Typically, distances or densities are used to give estimation what is normal and what is an outlier. This article only focuses on this unsupervised anomaly detection setup.

2 LITERATURE SURVEY

Data is an important asset to an organization or a company [13]. Organizing data can improve quality of data and be added value for the organization. One of the techniques that could be applied to ensure data quality is data profiling [14]. Data profiling is a process of examining the data available in a data source and collecting statistics and information of that data [15]. Data profiling is defined as the application of data analysis techniques to existing data stores for the purpose of determining the actual content, structure, and quality of the data. Data profiling is the set of activities and processes to determine the metadata about a given dataset [16].

Table 1: Comparative Analysis of Paper

Year of Publication	Title Of Paper	Description
2009	Anomaly Detection: A Survey	This paper provides structured and comprehensive overview of research on anomaly detection. It includes the definition, challenges, related work, various phases of anomaly detection problem, applications; several types of techniques etc. in short all about of anomaly detection [16].
2009	Detecting Anomalies in a Time Series Database	This paper presents a comprehensive evaluation of semi-supervised anomaly detection techniques for time series data. The techniques can be grouped into four categories, i.e., kernel, window, predictive, and segmentation- based techniques[16].
2016	Anomaly Detection In Aircraft Data Using Recurrent Neural Networks (RNN)	This paper describes the application of Recurrent Neural Networks (RNN) for effectively detecting anomalies in flight data [16].
2016	Anomaly based IDS using Backpropagation Neural Network	This paper presents the Anomaly Intrusion Detection System that can detect various network attacks. The goal is to identify those attacks with the support of supervised neural network that is. Back propagation artificial neural network algorithm and make complete data safe [16].

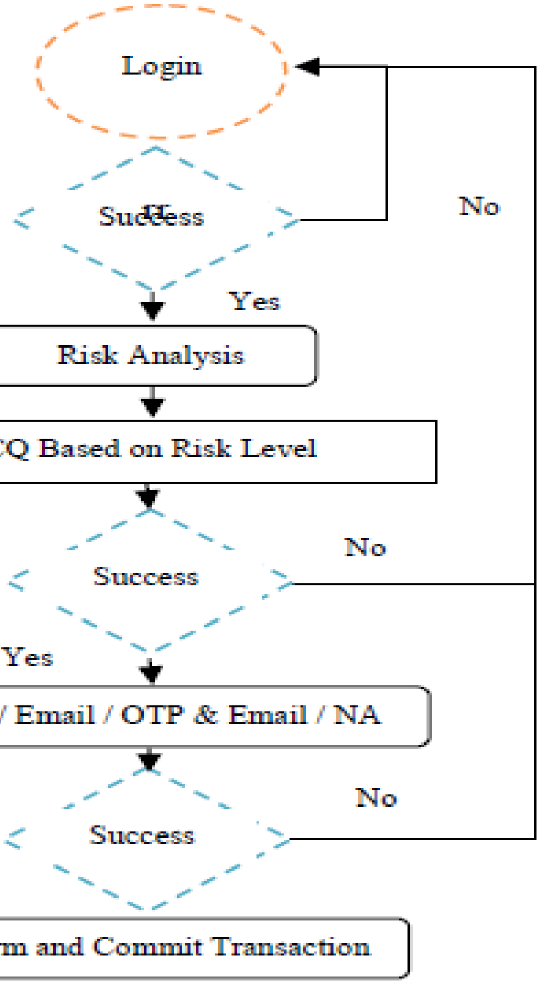
2016	Fuzzy Logic Inference for Unsupervised Anomaly Detection	This paper introduced the solution for unsupervised anomaly detection i.e., to detect unexpected activity of user or network equipment, based on the analysis of mutual dependencies of the separate slices of network activity [16].
2013	Anomaly Detection in Time Series Data using a Fuzzy C-Means Clustering	This paper presents anomalies in time series which are divided into two categories: amplitude anomalies and shape anomalies. A unified framework sustaining the detection of both types of anomalies is introduced [12].
2012	Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering	This paper work on to add restore point which allows for the rolling back of system files, registry keys, installed programs and the project data base [13].
2010	An Anomaly Detection Method Based on Fuzzy C-means Clustering Algorithm	This paper indicates the fuzzy C-means clustering (FCM) algorithm which applied to detect abnormality which based on network flow [14].

3 KNOW YOUR CUSTOMER

Internet based financial services like balance transfer, ecommerce transaction, bill payment, investment to bank products (saving certificate, fixed deposit revenue) etc. are the appealing ways of doing business as well as performing all financial transaction location independently. As a result of these overwhelming facilities, it poses the highest point of risk as it uses public network over the world. The main endeavor of the banks and non bank financial institutions are to provide a consistent, secured and high available process of authentication to their customers with minimizing potential avenues of attack, especially attacking vectors beyond the control of either the customers or the Financial Institutions (FIs).

In this model of authentication for financial application, some techniques of comprehensive risk factor judgment have, proposed to identify a suspicious login over many existing risk evaluation by analysis of user's historical activities data, and also proposed a new authentication method based on KYC information for authorizing the user during login as well as before performing transaction. In this method CQs are choosing from a collection of CQs where the level of the CQ is defined by the risk factors calculation result. Risk factors are calculated considering Credential Risk (New user, Failed login attempt, User with no/very few identity information, Changing nature of transaction), Behavioral Risk (Changed general transaction timing, Exceeded regular transaction frequency, changed regular transaction purpose, Called to a new URL, which did not call by this particular user previously), Transaction Risk (Exceeded regular transaction limit, Exceeded profile transaction limit), and Location Risk (Changed geo-location, Changed transaction device). For appropriate and secure transaction using application, proposed model performs two operations. First one is risk factor calculation and second one is assigning the CQ based on the risk assessment result.

The brief idea of the model is depicted in the Fig. 1. In the figure the initial step is login of the user with user ID and Password verification like other online applications. The forwarding stage is risk analysis for the login succeeded user. Next stage assign one or more CQ based on risk level formed by the result of prior stage. Final stage of the verification is OTP / EMAIL / OTP & Email confirmation if it is indicated by the result of risk analysis. In some other cases confirmation stage may not be applicable where CQ is in final stage before performing and committing a transaction. The CQ based authentication applies immediately after login and before performing and committing a transaction to verify the user rigorously. This CQ replaces the 2FA or traditional question and answers mechanism from some other existing authentication models [17].



4 CONCLUSION

As there is improvement in the technologies in the real world there are some of advantages and disadvantages in the technologies. Even the percentage of anomalies also been increased in the system. To overcome this, anomaly detection has been introduced with different techniques. Here the anomalies are identified in the KYC (know your customer) forms, where all the information of the customers are filled. There will be chances of customers with some of the details similar to the other customer so that could be some problem to identify the customer identity whether he/she is real customer in the provided details. Hence, anomalies are identified by using anomaly detection techniques.

REFERENCES

[1] Kumar, A. Banerjee, and V. Chandola, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, July 2009.

[2] Portnoy L, Eskin E, Stolfo S,” Intrusion Detection with Unlabeled Data Using Clustering”, In: In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001); 2001. p. 5–8.

[3] Garcia-Teodoro P, Diaz-Verdejo JE, Macia-Fernandez G, Vazquez E, “Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers and Security”, 2009;28:18 28.

[4] Yeung DY, Ding Y,”Host-Based Intrusion Detection Using Dynamic and Static Behavioral Models”, Pattern Recognition. 2003;36:229–243.

[5] Quinlan JR. C4.5: ,”Programs for Machine Learning”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.

[6] Schölkopf B, Smola AJ. “Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Adaptive Computation and Machine Learning”, MIT Press, Cambridge, MA; 2002.

[7] Mehrotra K, Mohan CK, Ranka S,”Elements of Artificial Neural Networks”, Cambridge, MA, USA: MIT Press; 1997.

[8] Moya MM, Hush DR”Network Constraints and Multi-objective Optimization for One-class Classification”, Neural Networks. 1996;9(3):463–474.

[9] Schölkopf B, Platt JC, Shawe-Taylor JC, Smola AJ, Williamson RC.,”Estimating the Support of a High-Dimensional Distribution. Neural Computation”, 2001;13(7):1443–1471. pmid:11440593

[10] Hawkins S, He H, Williams GJ, Baxter RA. “Outlier Detection Using Replicator Neural Networks”,In: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000). London, UK: Springer-Verlag; 2000. p. 170–180.

[11] Lindsay B,”Mixture Models: Theory, Geometry, and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics”, Penn. State University: Institute of Mathematical Statistics; 1995.

[12] Rosenblatt M,”Remarks on Some Nonparametric Estimates of a Density Function. The Annals of Mathematical Statistics”, 1956;27(3):832–837.

[13] Dorr and P. Herbert, "Data Profiling: Designing the Blueprint for Improved Data Quality," in SAS User Group International 30, Philadelphia, 2005.

[14] J. E.Olson, Data Quality The Accuracy Dimension, USA, 2013.

[15] Naumann, "Data Profiling Revisited," 2013.

[16] L. Golab, F. Naumann and A. Ziawasch, "Data Profiling," pp. 1-4, 2016.

[17] Prakash Chandra Mondal,et.al,” Know Your Customer (KYC) based authentication method for financial services through the internet”, 19th International Conference on Computer and Information Technology, 978-1-5090-4090-2/16/$31.00 ©2016 IEEE.