ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

Reverse Phishing for Anomaly Detection Using Clustering

R. Kumar, Saritha Maria Quinn, Vishnu Prasad

Department of Computer Science (PG), Kristu Jayanti College, Bengaluru, India

Department of Computer Science (PG), Kristu Jayanti College, Bengaluru, India

Department of Computer Science (PG), Kristu Jayanti College, Bengaluru, India

ABSTRACT - The latest germinating trend is plenary digitalization which implies that security; privacy and transparency in transactions need a secure platform and strategy to avoid forfeiture in financial assets. The era needs a more reliable method that examines traffic between the hosts, detects end-point attacks and sound network anomaly sniffing algorithm. Intrusions can be categorized into regular and irregular patterns. The existing authentication techniques are more vulnerable and fraudulent events like account take-over occurrences become more widespread. Bio-metrics has not paved its way in terms of materialization and cost-efficiency. Pattern-capturing can prove more efficacious since a plethora of attributes exist which record different user-patterns that could easily detect an anomaly. This pattern recognition is not a mere training of a model against normal and anomalous data. Security is enhanced by implementation of reverse phishing using Support Vector Machine. This implementation mainly focuses on using behaviour-based analysis of clients and financial institutions. Its performance lies in the fact that the algorithm focuses on behaviour patterns of both the user and the attacker. The frequent multivariate attributes are plotted and they do not just highlight attributes from a single perspective. The concept of reverse phishing that we propose is an intelligent strategy that inputs a trusted authentication from the party which demands the sensitive data.

KEYWORDS – Reverse Phishing; Support Vector Machine; User-patterns; Outliers.

I. INTRODUCTION

Data mining is the extraction of adequate relevant knowledge that can be statistically reliable in order to obtain potentially useful outcomes. It discovers relationship between two entities and can predict how objects would behave after being trained by a model. It also involves finding patterns and correlations to detect outliers. By adopting a wide range of analysis techniques, one could reduce business risks. Data mining has given a quantifiable approach to evaluate a model performance attribute. It could also give insights as to how best risks can be controlled.

Signature based anomaly detection is used only for that traffic that is recorded malicious is prevented from access. However, the drawback lies in the fact that new types of attack will not be identified. The ease of accessibility using pass-codes and the absence of a well-strategized transaction flow have resulted in an estimate of five cents per dollar being lost in fraud which could culvert the entire system. The various recognized fraudulent activities include Phishing, Pharming and injection of malicious spywares in order to extract user details. To overcome this, various Intrusion Detection Systems have evolved. Intrusion Detection Systems are provided to substantiate the security of communication and Information systems. In signature based detection, data is sniffed for known attacks rather than anomalous events while anomaly detection systems compare activities against a normal defined behavior. It is impossible to be absolutely certain about the legitimacy behind a transaction request. A more remunerative option is to extricate probable fraud threats from the available data using some logic algorithms. This is exactly what signature based systems detect. Basically, they act on known data streams. This area has become one of the most established fields in the context of data mining applications. They find application in sensitive areas such as Credit card or identity thefts.

Another detection technique is Anomaly based detection. This technique builds a baseline of what’s normal and the segregates abnormal data packets/requests. Clustering finds application in this implementation as if it were not for this technique; anomaly detection would have been complex. It employs the Support Vector Machine [6]as it uses a kernel algorithm for pattern recognition.

This paper highlights several anomaly detection schemes for identifying normal and anomalies. It is organized as follows. Section II conveys the existing difficulties and solution for anomaly detection. Section III portrays the implementation of Support Vector Clustering algorithm. Section IV concludes with future enhancement.

II. RELATED WORK

The method of cascading k-means clustering and ID3 decision tree proposed by K. Hanumantha Rao [1] implemented anomaly detection using machine-learning. Shekhar R. Gaddam, Vir V. Phoha, and Kiran S. Balagani [7] adopted a K-Means and ID3. A Novel Method for Supervised Anomaly Detection achieved by Cascading K-Means Clustering. But, the system will not be defensive enough for a new attack. The main disadvantages of the existing systems are false positives and false negatives. The major drawback in the existing system lies in the inability to authenticate the service that carries and requests the sensitive data. The existing methods moreover prove inefficient when encountered by gradual change in the data flows. [5] ZacharyMiller, William Dietrich and Wei Huhave performed a significant survey on the implementation of different IDS strategies. Many expert systems were analyzed. NIDES ALADand PHAD were developed in order to monitor the network based on a signature model. They could detect the IP address, packet headers and port numbers. But, they succumb to byte padding and substitution mechanisms [2] from the attacker. NETAD came out with creating different network modules for each network protocol which proved more effective. And the last technique which looks close to the proposed system is the behavior based anomaly detection. But, what makes this our proposed system different is that it implicates a definite series of logical attributes that cannot be detoured or skipped.

We propose a more reliable system where the embedded IDS algorithm to detect encrypted packets, scrambled source addresses, and to exactly predict possible attacks. While the users are educated about creating unique, customized patterns, the unauthorized user trying to get hold of the account after being detected by an intelligent system, gets tricked by their outlier user-pattern and the account would be temporarily blocked until the user re-types a security pattern onto the system via a virtual keypad in order to avoid key-logging. This eliminates the complex procedure of resetting the account. A session time-out while not receiving user pattern can also be termed as an outlier. This process of portraying deceptive error messages in order to defend against deception could be termed as “Reverse-Phishing.”Moreover, it implicates “user-patterns” rather than “transaction patterns”

III. DATA PREPARATION

The data-source for the analysis would be network traffic and user-patterns which could be customized or understood inherently by the system over a period. Network granular information like web-domains and URL’s could be captured as our data source. Customizable user-patterns could also be instantiated using personal information given by the customer [4]during the time of account creation. Security patterns must be stored and kept confidential by the user.

‘

IV. METHODOLOGY

T here are a plethora of parameters that explain correlation of attributes like clusters, principal components, correlations or classifications. Data in raw representations should be transformed and pre-processed into meaningful representations. Z score normalization equation 1, also called as Zero mean normalization is the sum of the all attribute values divided by the standard deviation.

$$d' = \frac{d - mean(P)}{std(p)} \text{----} \ (1)$$

The system we proposed adds performance and weightage to the existing one by adding an efficient IDS algorithm which can be embedded in a browser and activated while the user activates a network connection. This saves memory and process time.

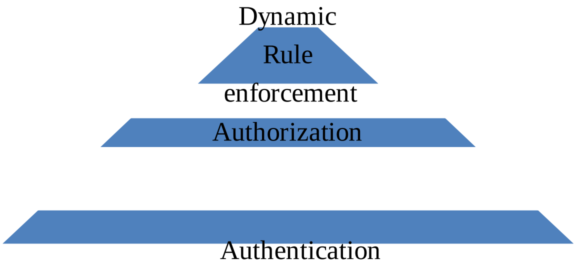


Fig. 1 Visibility and control of Security Architecture

Authentication is performed for both perspectives before user transmits his sensitive data so that transaction is protected and the latter must provide access to the authentic user.

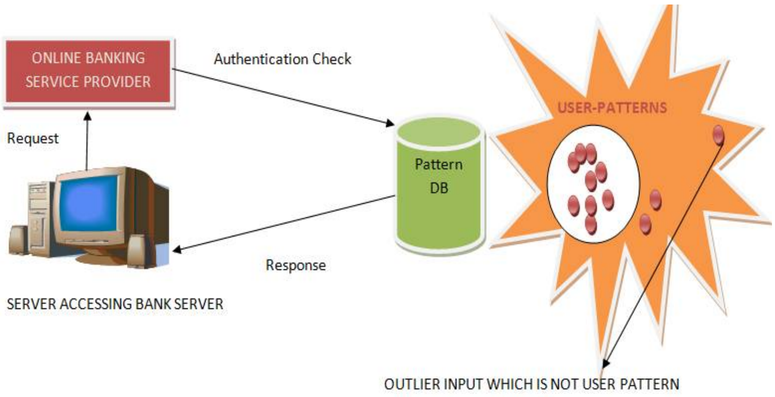


Fig 2.Illustration of interaction between modules

Algorithm for anomaly check:

```
Step 1: Provide input for analysis
Step 2: Normalize the data using Z-normalization.
Step 3: Apply support vector clustering to the normalized data
Step 4: if (data==anomaly)
Step 5: then →input encrypted pattern
Step 6: if (pattern=anomalous)
Step 7: Block transaction until security pattern is entered;
Step 8: else
Step 9: Complete transaction
```

It has been experimentally proven that z-score normalization is more accurate when compared to min-max normalization [6]. It also preserves data privacy. The data attributes in each personalized cluster are conjoined using a kernel function. The smallest sphere that encloses is selected and using the Support Vector Domain Description algorithm. The attribute space will be the smallest sphere that encloses the data-image. The number of incoherent contours is inversely proportional to the distance of the kernel [8]. Hence it deals more precisely with outliers by employing a threshold that allows the sphere in attribute-space to not to enclose all points[9].The cluster objects clustering is shown in detail in Fig 3 and Fig 4. The noise ratio is also depicted graphically.

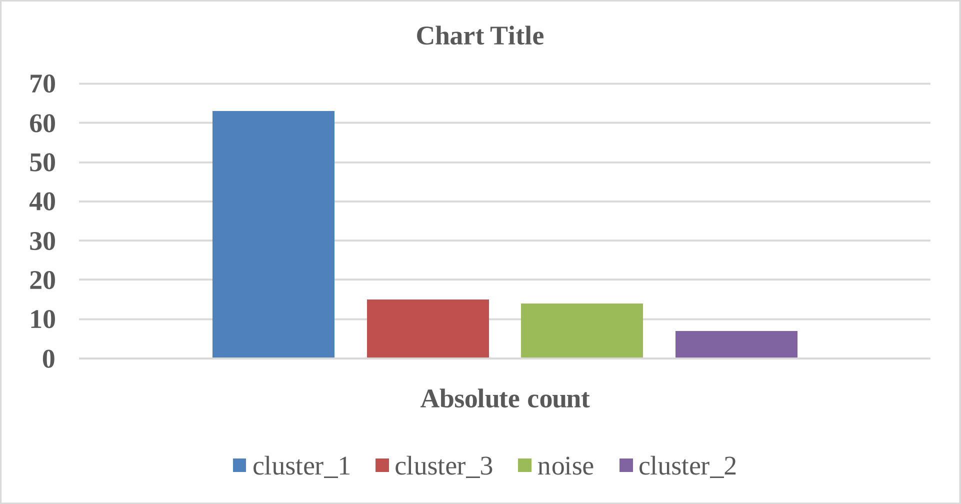
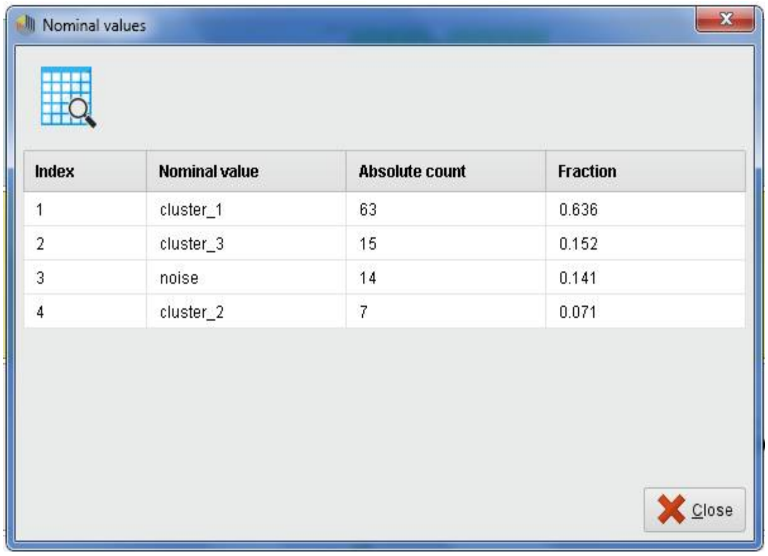


Fig 3. Support Vector Clustering



Index	Nominal value	Absolute count	Fraction
1	cluster_1	63	0.636
2	cluster_3	15	0.152
3	noise	14	0.141
4	cluster_2	7	0.071

Fig 4. Nominal values of Clustering for the sample data set

The following flowchart in Fig. 5 represents the working of the proposed system.

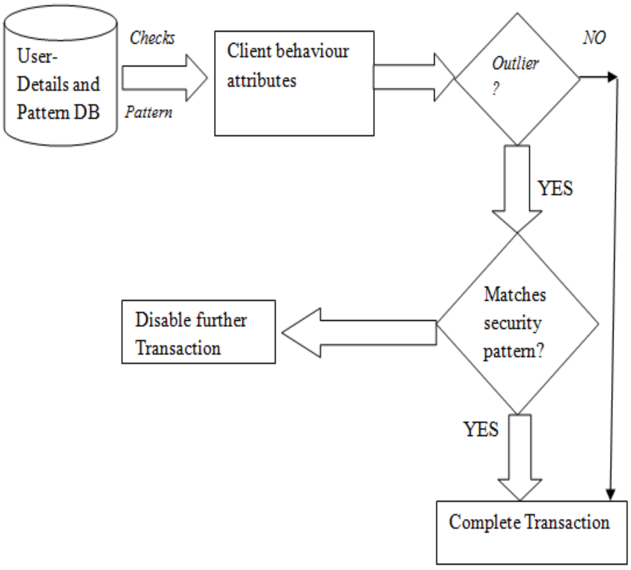

User-Details and Pattern DB
Checks
Pattern
Client behaviour attributes
Outlier ?
NO
YES
Matches security pattern?
Disable further Transaction
YES
Complete Transaction

Fig 5. System Implementation flow diagram

Method	No. of frauds detected	False-positives	False Negatives	Accuracy
Supervised	50	10	3	74%
Unsupervised	50	7	2	82%

Table 1 Comparative Performance for various techniques

V. CONCLUSION

Traditional defense strategies cannot keep-up with the current threats. A feature aimed at providing improved customization and alternatives become a loop-hole to online fraudsters. Advanced malware and polymorphic blending could bypass primary and secondary authentications. We could now integrate the proposed strategy of pattern recognition and a Network intrusion detection algorithm like NIDES which can confirm the malicious and fraudulent act by using reverse phishing. The technique of detecting anomalies in networks packets has already been researched upon. In order to make NIDES more effacious, we need an unsupervised method of detecting new anomalous attributes and not just blocking mimic attacks. Rapid resolution becomes necessary in order to safeguard transactions involving sensitive data. The proposal shows precision and accuracy levels 84% being the minimum. Hence it would be suitable for security applications.

REFERENCES

[1] K. Hanumantha Rao, G. Srinivas, AnkamDamodhar and M. Vikas Krishna:"Intrusion Detection System using Data mining Techniques", International Journal of Computer Science and Telecommunications ,Volume II, Issue 3, June 2011.

[2] W. Lee, S. J,"Data Mining Approaches for Intrusion Detection" Proceedings of the 7th USENIX Security Symposium, January 26-29, 1998, San Antonio, Texas.

[3] Rui Xu,, Donald Wunsch II, "Survey of Clustering Algorithms", IEEE in Neural Networks 16(3) (2005)

[4] R.Perdisci, "Statistical Pattern Recognition Techniques for Intrusion Detection in Computer Networks, Challenges and Solutions," University of Cagliari, Italy, 2006. M. Mahoney and Models of Normal Network Traffic for Detecting Novel Attacks,"

[5] Zachary Miller, William Deitrick, Wei Hu* "Anomalous Network Packet Detection Using Data Stream Mining", IEEE Transactions on Neural Networks,Volume: 16, Issue: 3, May 2005

[6] Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," Proc. SIAM Int'l Conf. Data Mining, May 2003.

[7] Shekhar R. Gaddam, Vir V. Phoha, and Kiran S. Balagani ," K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods", IEEE Transactions on Knowledge and Data Engineering, VOL. 19, NO. 3 March 2007.

[8] Karl-Heinrich Anders," A Hierarchical Graph-Clustering. IEEE Transactions on Approach to find Groups of Objects

[9] Macgregor, M.Hall, P.Lorier and J.Bruskill, "Flow clustering using machine learning techniques", In PAM 2004, Antibes-Juan-Les-Pins, France, LNCS. pp. 205-214, 2004.

BIOGRAPHY

R. Kumar is Associate Professor and Head, Department of Computer (PG) Kristu Jayanti Collegem Bangalore. He received Master of Computer r Science degree in 1992 from Bharathidasan University, Trichy, South India MS, India. His research interests are Data Mining, Motif Discovery. Pursuing Ph.D. in MS University, Tirunelveli, South India.

Saritha Maria Quinn and Vishnu Prasad are pursuing MCA, Kristu Jayanti College, Bangalore. Their area of research interest are datamining and network security.