Anomaly detection for weblog data analysis using weighted PCA technique

Article *in* Journal of Information and Optimization Sciences · January 2022

DOI: 10.1080/02522667.2022.2037283

2 authors, including:

Suman Mann
Maharaja Surajmal Institute Of Technology
71 PUBLICATIONS **248** CITATIONS

SEE PROFILE

Journal of Information and Optimization Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tios20

Anomaly detection for weblog data analysis using weighted PCA technique

Meena Siwach & Suman Mann

To cite this article: Meena Siwach & Suman Mann (2022) Anomaly detection for weblog data analysis using weighted PCA technique, Journal of Information and Optimization Sciences, 43:1, 131-141, DOI: 10.1080/02522667.2022.2037283

To link to this article: https://doi.org/10.1080/02522667.2022.2037283

Published online: 28 Mar 2022.

Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 43 (2022), No. 1, pp. 131–141
DOI : 10.1080/02522667.2022.2037283

Taylor & Francis
Taylor & Francis Group

Anomaly detection for weblog data analysis using weighted PCA technique

Meena Siwach *
Guru Gobind Singh Indraprastha University
India

and

Department of Information Technology
Maharaja Surajmal Institute of Technology,
Delhi 110078
India

Suman Mann †
Department of Information Technology
Maharaja Surajmal Institute of Technology
Delhi 110078
India

Abstract

Many methods have been developed to protect web servers against attacks. Anomaly detection methods rely on generic user models and application behavior, which interpret departures as indications of potentially dangerous behavior from the established pattern. However, due to a lack of evaluations and comparisons of various anomaly detection techniques, engineers may still decide which detection methods should not be used. Furthermore, even if engineers use an unusual detection technique, re-implementation will take a lifetime. We offer a comprehensive analysis and evaluation of six existing log-based detection techniques, including three monitored and three unchecked modes, as well as an open toolkit that allows for simple reuse, to address these problems. The different anomalies are detected with weighted PCA techniques. There are four datasets BGL, Liberty, Spirit

* *E-mail:* meenusiwach@gmail.com (Corresponding Author)
† *E-mail:* sumanmann2007@gmail.com

& Thunderbird, which are used. The weighted PCA is compared with traditional KNN methods. The weighted PCA provided better results as compared to the KNN algorithm.

Subject Classification: *68xx, 90xx.*

Keywords: *Weighted PCA, BGL, Liberty, Accuracy, Precision rate.*

1. Introduction

Due to their high value, Web servers are gradually becoming targets for assaults as the information technology sector advances. SQL injection and cross-site scripting (XSS) threats have been increasingly common in recent years, which is why Web security has received more attention from academic and industry communities. Anomaly is a term used in internet security research. The analysis of log data is used in web detection [1]. Log files, as crucial recording data, may reveal extensive information at the time of system operation and may be used to trace the majority of assaults. However, log systems create a lot of data, and critical information might be lost in the shuffle [2, 3].

In addition, conventional intrusion detection techniques involve operators or programmers to remove the attack features manually, and detect common attack patterns depending on searching of keyword and rule matching [4]. In other words, the conventional approach cannot detect unidentified attacks and leads to failure. A variety of anomaly detection techniques are being suggested to solve the limitations of conventional methods to overcome the shortcomings of previous years. Many machine learning methods are being utilized in the identification of log-based anomalies as a result of advances in machine learning [5]. Anomaly detection techniques are typically split into two groups based on the kind of data and the use of machine learning technology: supervised detection [6] and unsupervised detection [7,8]. Normal training data, properly described in both good and bad circumstances, is required for the supervised approach. Unsupervised techniques, on the other hand, do not need labels at all. Their research is based on the fact that a traumatic experience may sometimes act like a far-fetched advertisement [9,10].

Luo et al. proposed architecture for detecting anomalies in log data that does not need any former knowledge of the domain. The proposed method includes a process for diagnosing log key and parameter value abnormalities, as well as a mechanism for identifying log key and

parameter value abnormalities from logs. The probability of the next log key is predicted using a neural network-based method [11].

A log parameter sequence abnormality can similarly be detected using a comparable LSTM neural network. The software also uses false-positive manual feedback to improve future accuracy. The LSTM considers the log series to be a natural language sequence that may be processed accordingly. Using datasets from BGL, Thunderbird, Open Stack, and IMDB [12]. Guangmin et al. suggested a deep learning model for detecting log message abnormalities and compared these models to boost efficiency [13]. The IMDB dataset is used to demonstrate how their method can be used to a range of classification challenges [14].

Natural Language Processing techniques were used by Zolotukhin et al. [15] to discover abnormal log messages. In the research, word2vec and TF-IDF feature extraction methods are applied, and the activity is finished with a classification LSTM deep learning algorithm. They discovered that word2vec beats TF-IDF in log message identification jobs [16].

In this paper, a weighted PCA-based system of anomaly detection is proposed for the Weblog file. To reduce the above-mentioned disadvantages of the traditional method, this system uses an algorithm of weighted PCA on two levels. The decision tree classifier is used to choose standard log files, and then Markov's hidden model is used to build a standard data model set (hereafter HMM) [16]. This is an example of a model for detecting anomalies. It generates automated knowledge and training based on a large amount of data, raising the stakes in the online security debate to new heights.

The paper is organized as follows. In section 2, we discussed past research work. Section 3 discusses the proposed methodology for anomaly detection. Section 4 provides the result analysis of the proposed work. Finally, in section 5 we conclude and consider future work.

2. Proposed Methodology

In this section, a novel weighted PCA algorithm is discussed. The proposed methodology framework is discussed in figure 1. The anomaly detection of weblog data was analyzed with the weighted PCA method [10]. The particular framework consists of log data vectorization, labeling, and anomaly detection.

3.1 *Log Parsing Method*

Logs are simple texts with elements that may vary from one instance to the next. The words "Connection from 10.10.34.12 closed" and "Connection from 10.10.34.13 closed" are considered constant parts in the logs "Connection from 10.10.34.12 closed" and "Connection from 10.10.34.13 closed," for example, because they never change, but the remaining parts are known as variable parts because they are not fixed [17]. Although developers specify constant components in source codes, variable portions (such as port numbers and IP addresses) are sometimes dynamically generated and hence unsuitable for anomaly detection [18]. The purpose of log parsing is to extract constants from variable items and generate a well-defined log event (for example, "Connection from * ended"). Both cluster-based. Clustering-based log parsers determine the distances between logs in the first phase, and clustering algorithms are then used to organize the logs into discrete groups in the second phase. Each cluster generates a template for an event [19].

In heuristic-based approaches, the number of times each word appears in each log position. Common words are chosen and produced as event candidates. Last but not least, decide which candidates will be registered as events [20]. In pre-work, we created and compared four log parsers [21]. We also made an open-source log parsing toolkit available online, which we used to parse raw logs into log events for our research.

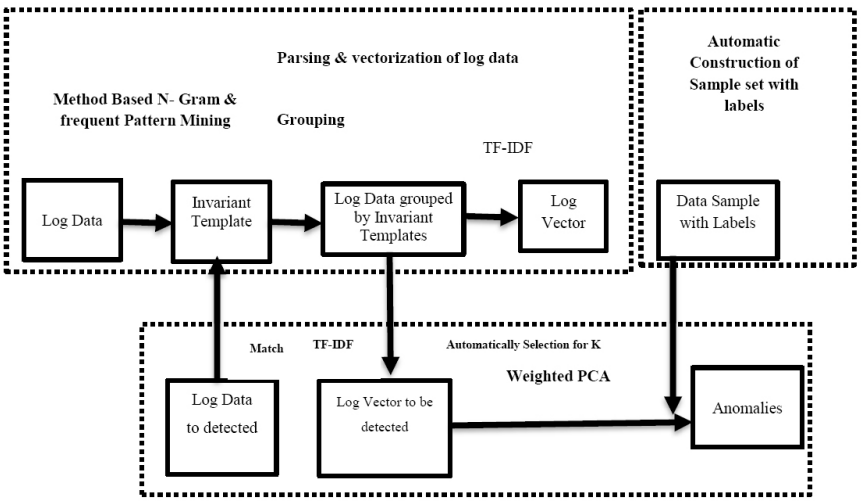

Fig. 1

The overall framework of Proposed Technique

3.2 *Log- Data Vectorization*

The different data are categories into group data form as log lines represented by l_j having words w_i with magnitude is represented by [22]

$$W(w_i, l_j) = tf(w_i, l_j) * \log\left(\frac{N}{n_{w_i}} + 0.1\right) \quad (1)$$

In this, the different weights for the appropriate word $w_i$are represented by $W(w_i, l_j)$. & Sample size is represented by N.

3.3 *Anomaly Detection with the Weighted PCA*

The anomaly detection with weighted PCA follows the three steps as weblog data enhancement [23,24]. And the log data vectorization. The proposed weighted PCA model is represented by

$$u_{mx_1} = W_{mx_d} xd_{x_1} \quad (2)$$

Where u, an m-dimensional vector, is a projection of x - the original d-dimensional data vector ($m << d$).

3. Result Analysis

The performance parameter of weblog datasets is shown in table 1. In this table, four datasets BGL, Liberty, Spirit & thunderbird are used. The maximum recall rate & accuracy achieved for the BGL dataset is 100 % & 97.62 % respectively. Similarly, the maximum F1-score achieved for the Spirit dataset is 98.19 %

Table 1

Performance Parameter of Web Log Datasets

Sr. No.	Dataset	Recall Rate	F 1-Score	Accuracy
1	BGL	100 %	96.55 %	97.62 %
2	Liberty	96.29 %	94.52 %	92.83 %
3	Spirit	98.79 %	98.19 %	97.60 %
4	Thunderbird	97.53 %	96.34 %	95.17 %

Table 2
Comparative analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set

Sr. No.	Parameter	I-KNN	I-PCA Ensemble
1	Accuracy	91.73 %	**97.62 %**
2	Recall Rate	90.97 %	**100 %**
3	F-1 Score	92.784 %	**96.55 %**

Table 3
Comparative analysis of I-KNN & I-PCA Ensemble on Spirt/2 Log set

Sr. No.	Parameter	I-KNN	I-PCA Ensemble
1	Accuracy	91.97 %	**97.60%**
2	Recall Rate	**99.15 %**	98.79 %
3	F-1 Score	96.13 %	**98.19 %**

Table 4
Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets.

Sr. No.	Data Set	I-KNN	I-PCA Ensemble
1	BGL	92.15 %	**97.62%**
2	Liberty	92.32 %	**92.83 %**
3	Spirit	91.14 %	**97.60 %**
4.	Thunderbird	**95.27 %**	95.17 %

Table 2 and Table 3 represent the comparative analysis of the I-KNN & I-PCA Ensemble on the BGL/2 Log set and Spirt/2 Log set. In the case of the BGL/2 Log Set, the accuracy, recall rate, and F1-Score for Weighted KNN is 91.73 %, 90.97 %, and 92.784 % respectively. Similarly, the accuracy, recall rate, and F1-Score for the Weighted PCA Ensemble technique are 97.62 %, 100 %, and 96.55 % respectively. In the case of Spirit/2 Log Set, the accuracy, recall rate, and F1-Score for Weighted KNN is 91.97 %, 99.15

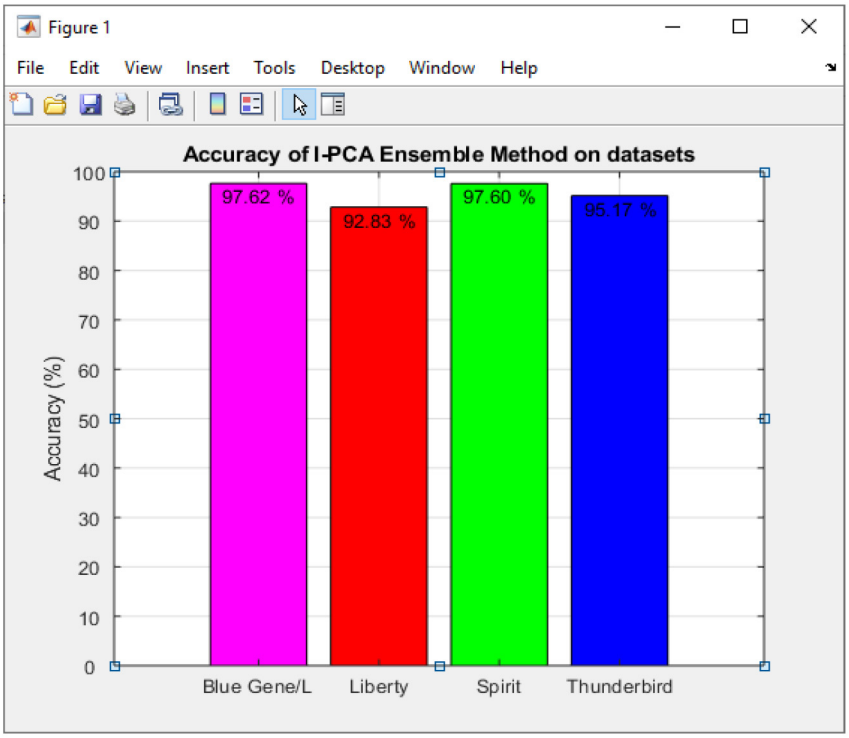

Fig. 2

Computation of Accuracy of Weighted PCA Ensemble Method on different Datasets

%, and 96.13 % respectively. Similarly, the accuracy, recall rate, and F1-Score for the Weighted PCA Ensemble technique are 97.60 %, 98.79 %, and 98.19 % respectively.

Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets is represented in table 4. The accuracy for BGL, liberty, Spirit & thunderbird datasets is 92.15 %, 92.32 %, 91.14 % and 95.27 % respectively. Similarly, the accuracy for BGL, liberty, and Spirit & thunderbird datasets is 97.62 %, 92.83 %, 97.60 %, and 95.17 % respectively. The maximum accuracy achieved with the weighted PCA Ensemble algorithm for BGL, Liberty, and Spirit is 97.62 %, 92.83 %, and 97.60 % respectively. Similarly, the maximum accuracy achieved with the weighted KNN algorithm for thunderbird is 95.27 %.

Comparative Analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set it represents in fig 3. The accuracy, recall rate, and F1-Score for

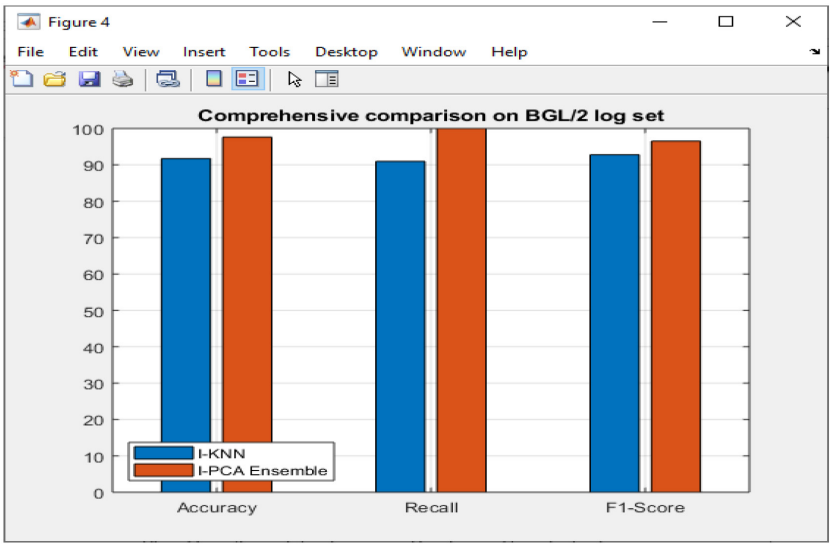

Fig. 3

Comprehensive comparative analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set

Weighted KNN are 91.73 %, 90.97 %, and 92.784 % respectively. Similarly, the accuracy, recall rate, and F1-Score for the Weighted PCA Ensemble technique are 97.62 %, 100 %, and 96.55 % respectively. It represents that the performance parameter of the BGL/2 Log set is enhanced by the weighted PCA Technique.

4. Discussion and Conclusion

In modern large systems, **the main objective of our research is to design novel techniques used for anomaly detection. One of the primary limitations identified in this systematic review is the absence of a standardized, up-to-date, and properly labeled dataset that enables the verification of experimental results acquired in various investigations.** The overall optimized result is obtained from the weighted PCA algorithm. The maximum recall rate & accuracy achieved for the BGL dataset is 100 % & 97.62 % respectively. Similarly, the maximum F1-score achieved for the Spirit dataset is 98.19 %. The accuracy, recall rate, and F1-Score for the Weighted PCA Ensemble technique are 97.62 %, 100 %, and 96.55 % for BGL/2 Log Set Data. Similarly, the accuracy, recall rate, and F1-Score for

the Weighted PCA Ensemble technique are 97.60 %, 98.79 %, and 98.19 % respectively for Spirit/2 log set data.

References

[1] Zhao, Zhijun, Chen Xu, and Bo Li. "An LSTM-Based Anomaly Detection Model for Log Analysis." *Journal of Signal Processing Systems* 93, no. 7 (2021): 745-751.

[2] Ieracitano, Cosimo, Ahsan Adeel, Francesco Carlo Morabito, and Amir Hussain. "A novel statistical analysis and autoencoder driven intelligent intrusion detection approach." *Neurocomputing* 387 (2020): 51-62.

[3] Fernandes, Gilberto, Joel JPC Rodrigues, Luiz Fernando Carvalho, Jalal F. Al-Muhtadi, and Mario Lemes Proença. "A comprehensive survey on network anomaly detection." *Telecommunication Systems* 70, no. 3 (2019): 447-489.

[4] Kwon, Donghwoon, Hyunjoo Kim, Jinoh Kim, Sang C. Suh, Ikkyun Kim, and Kuinam J. Kim. "A survey of deep learning-based network anomaly detection." *Cluster Computing* 22, no. 1 (2019): 949-961.

[5] Xiao, Ruliang, Jiawei Su, Xin Du, Jianmin Jiang, Xinhong Lin, and Li Lin. "SFAD: Toward effective anomaly detection based on session feature similarity." *Knowledge-Based Systems* 165 (2019): 149-156.

[6] Luo, Yuxuan, Shaoyin Cheng, Chong Liu, and Fan Jiang. "PU learning in payload-based Web anomaly detection." In *2018 Third International Conference on Security of Smart Cities, Industrial Control System, and Communications (SSIC)*, pp. 1-5. IEEE, 2018.

[7] Vartouni, Ali Moradi, Saeed Sedighian Kashi, and Mohammad Teshnehlab. "An anomaly detection method to detect web attacks using stacked auto-encoder." In *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pp. 131-134. IEEE, 2018.

[8] Najafabadi, Maryam M., Taghi M. Khoshgoftaar, Chad Calvert, and Clifford Kemp. "User behavior anomaly detection for application-layer DDoS attacks." In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 154-161. IEEE, 2017.

[9] Zolotukhin, Mikhail, Timo Hämäläinen, Tero Kokkonen, and Jarmo Siltanen. "Increasing web service availability by detecting application-layer DDoS attacks in encrypted traffic." In *2016 23rd International conference on telecommunications (ICT)*, pp. 1-6. IEEE, 2016.

[10] Ren, Xin, Yupeng Hu, Wenxin Kuang, and Mohamadou Ballo Souleymanou. "A Web Attack Detection Technology Based on Bag of Words and Hidden Markov Model." In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 526-531. IEEE, 2018.

[11] Kozik, Rafał, Michał Choraś, and Witold Hołubowicz. "Hardening Web Applications against SQL Injection Attacks Using Anomaly Detection Approach." In *Image Processing & Communications Challenges 6*, pp. 285-292. Springer, Cham, 2015.

[12] Valeur, Fredrik, Giovanni Vigna, Christopher Kruegel, and Engin Kirda. "An anomaly-driven reverse proxy for web applications." In *Proceedings of the 2006 ACM symposium on Applied computing*, pp. 361-368. 2006.

[13] Guangmin, Liang. "Modeling unknown web attacks in network anomaly detection." In *2008 Third International Conference on Convergence and Hybrid Information Technology*, vol. 2, pp. 112-116. IEEE, 2008.

[14] Sakib, Muhammad N., and Chin-Tser Huang. "Using anomaly detection based techniques to detect HTTP-based botnet C&C traffic." In *2016 IEEE international conference on communications (ICC)*, pp. 1-6. IEEE, 2016.

[15] Zolotukhin, Mikhail, Timo Hämäläinen, Tero Kokkonen, and Jarmo Siltanen. "Analysis of HTTP requests for anomaly detection of web attacks." In *2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing*, pp. 406-411. IEEE, 2014.

[16] Zhang, Ming, Shuaibing Lu, and Boyi Xu. "An anomaly detection method based on multi-models to detect web attacks." In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, pp. 404-409. IEEE, 2017.

[17] Parhizkar, Elham, and Mahdi Abadi. "OC-WAD: A one-class classifier ensemble approach for anomaly detection in web traffic." In *2015 23rd Iranian Conference on Electrical Engineering*, pp. 631-636. IEEE, 2015.

[18] Kozik, Rafal, and Michal Choras. "Adapting an ensemble of one-class classifiers for a web-layer anomaly detection system." In *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, pp. 724-729. IEEE, 2015.

[19] Cao, Qimin, Yinrong Qiao, and Zhong Lyu. "Machine learning to detect anomalies in weblog analysis." In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 519-523. IEEE, 2017.

[20] Hooda, S., and S. Mann. "A Focus on the ICU's Mortality Prediction Using a CNN-LSTM Model." International Journal of Psychosocial Rehabilitation 24, no. 6 (2020): 8045-8050.

[21] Sakshi Hooda, Suman Mann, "Sepsis-Diagnosed Patients' In-Hospital Mortality Prediction Using Machine Learning: The Use Of Local Big Data-Driven Technique in the Emergency Department" International journal of the grid and distributed computing, vol13, Issue 1 2020

[22] Anish Batra, Guneet Singh Sethi, Suman Mann," Personalized Automation of Electrical and Electronic Devices Using Sensors and Artificial Intelligence—the Intelligizer System" Computational Intelligence: Theories, Applications and Future Directions - Volume I, 2019, Volume 798

[23] Ruchika Kaushik, Vijander Singh, Rajani Kumar, "Multiclass-class SVM based network intrusion detection with attribute selection using infinite feature selection technique" *Journal of Discrete Mathematical Science and Cryptography*, 24, no. 8 (2021) : 2137-2153.

[24] Kumar, Ankit, et al. "An enhanced quantum key distribution protocol for security authentication." *Journal of Discrete Mathematical Sciences and Cryptography* 22.4 (2019): 499-507.