International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING

ISWARYA.G[1], PREVEENA.T[2],OVIYA.M.S[3],Dr.K.L.NEELA[4]

[1,2,3] B.E., Final Year Students, [4]Assistant Professor

Department of Computer Science and Engineering-University College of Engineering Thirukkuvalai

(A Constituent College of Anna University :: Chennai and Approved by AICTE, New Delhi)

ABSTRACT

The proposed scheme works five predictive analytics are utilizing assessed various metrics with an outsized collection of inputs values sets each model computes a spam score by considering the refined input characters this achieve depicts the reliability of IOT device under various parameters refit home dataset is employed for the validation of suggested technique the results achieved proves the efficacious of the proposed scheme as compared to the opposite existing schemes the amount of knowledge discharged from these devices will increase many-fold within the years to return additionally to an enlargement in volume the produces an oversized amount of information with variety of various modalities having varying data quality defined by its expedition in response your interval position dependency in such an environment predictive analytics mathematic could be play a significant role in ensuring sanctuary and authorization supported biotechnology anomalous detection to boost the usability and security of IOT systems on the opposite hand attackers often view analytic mathematic achievement feat the susceptibilities in canny systems motivated from these during the project we propose the protection of the IOT devices by detecting ssspam using machine learning to realize this objective spam detection in IOT utilizing predictive analytic armature is intentioned

KEYWORDS

Spam Detection, Machine Learning, Support Vector Machine(SVM), Multilayer Perceptron (MLP), K-nearest neighbor(KNN),Light GBM

1. INTRODUCTION

Internet of Things (IOT) give authority combination and accomplishment between the real-world objects disregarding of their geographical locations. Implementation of such network management and control make seclusion and safeguard strategies utmost important and challenging in such an environment. IOT applications need to protect data seclusion to fix security issues such as intrusions, spoofing attacks, DOS attacks, and DOS attacks, jamming, eavesdropping, spam, and malware.

It is common at the workplace that the IOT devices installed in an organization can be used to implement security and privacy attributes efficiently. For example, multifunctional appliance collect and send user's health data to a connected smart phone should prevent leakage of information to ensure privacy. It has been found in the market that 25-30% of working employees connect their personal IOT devices with the organizational network. The expanding nature of IOT attracts both the audience, i.e., the users and the attackers

1. SYSTEM ARCHITECTURE

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture Fig2.1.1 description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.

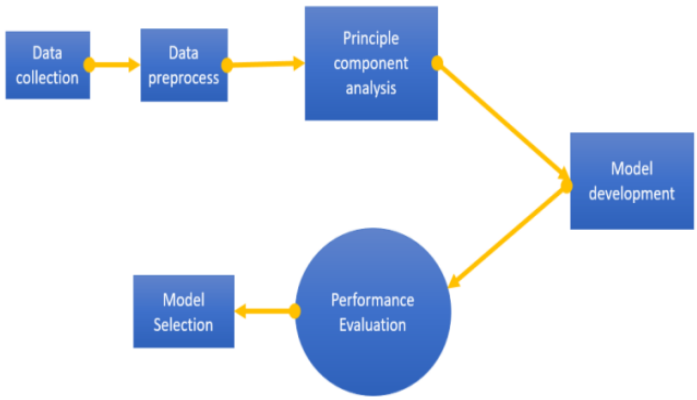

Fig.1. Architecture Diagram

2. MODULE LIST

- Data Collection
- Data preprocessing
- PCA Analysis
- Model Development
- Performance Evaluation
- Model selection

2.1. DATA COLLECTION

In this process to collect a data from Kaggle. The website like https://www.kaggle.com. This Kaggle data collection has an only numerical value in REFIT data sets https://www.refitsmarthomes.org/datasets and in this data using multiple purpose. Kaggle supports a variety of dataset publication formats, but we strongly encourage dataset publishers to share their data in an accessible, non-proprietary format if possible.

2.2. DATAPREPROCESSING

- Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

- A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning mode.

2.3. PCA ANALYSIS

- In this process collecting a preprocessed data to convert to model development process.

2.4. MODEL DEVELOPMENT

- We are using three different Algorithm in this model development process.

1. KNN
2. SVM
3. Light BGM

3.5.PERFORMANCE EVALUATION

In this process to implement of project Accuracy, recall and Precision

3.5.1 ACCURACY

Accuracy is one metric for evaluating classification models. Machine learning model accuracy is that the measurement accustomed to determine which model is best at identifying relationships and patterns between variables in an exceedingly dataset supports the input, or training, data. Accuracy is defined because the percentage of

correct predictions for the test data. It is often calculated

easily by dividing the quantity of correct predictions by the amount of total predictions.

$$accuracy = \frac{\text{correct predictions}}{\text{all predictions}}$$

3.5.2. PRECISION

Precision is defined as the fraction of relevant instances among the retrieved instances. In simple words, it is the ratio between actuality positives and each one positives. Precision helps when the prices of false positives are high.

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

3.5.3. RECALL

It's the amount of correct positive results divided by the quantity of all relevant samples. High recall means an algorithm returned most of the relevant results.

$$recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

ALGORITHM	ACCURACY
Support Vector machine (SVM)	88%
Multi Layer Preceptron	99%
Light GBM	94%

Fig 2:Performance Metrics table

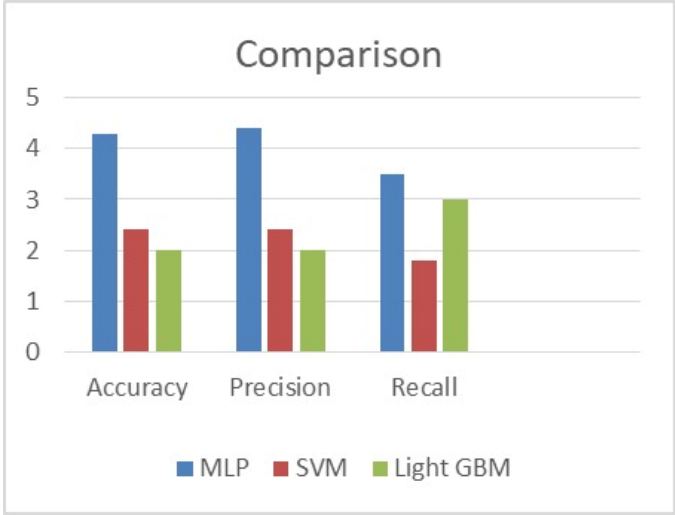

Fig 3: Performance Metrics Comparison

3.4. MODEL SELECTION

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset

4. MACHINE LEARNING TECHNIQUE

The use of machine learning models within the IOT has shown promising results for identifying malicious internet traffic using anomaly detection research. Moreover, either detection of anomalies or the employement of a spamicity score to trace the safety of the network components are motivated to possess a safe and secure network infrastructure. Several ML models are utilized for supervised machine learning; however, this paper uses ensemble methods, a group of ML techniques supported by decision trees. The machine learning models utilized within the paper are described as follows.

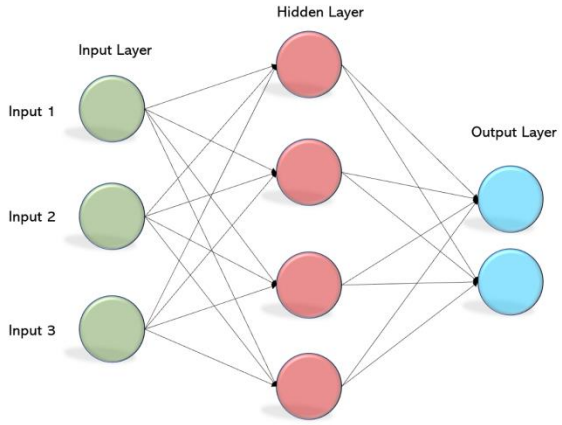

Fig 4: Multi Layer Perceptron

4.1.1.SUPPORT VECTOR MACHINE(SVM)

Support vector machines, also referred to as support vector networks, are a group of related supervised learning methods used for classification and regression. However, it's mostly utilized in classification problems. Within the SVM algorithm, we plot each data item as a degree in n-dimensional space (where n is the number of features you have) with the worth of every feature being the worth of a specific coordinate. Then, we can classify by finding the hyper-plane that differentiates both classes. Hence, we can say that the main objective of SVM is to finda hyperplane in an N- dimensional space that distinctly classifies the data points. SVM can classify both linear and non-linear data. To classify non-linear data it uses a method called the kernel trick to rework your data so it supports these transformations and it also finds an optimal boundary between possible outputs. A kernel is a function which maps a lower dimensional data into higher dimensional data. Simply put, it does some extremely complex data transformations, then figures out a way to separate your data. supporting the labels or outputs you've defined. Given a group of coaching examples, each marked as belonging to 1 of two categories, an SVM training algorithm builds a model that predicts whether a replacement example falls into one category or the opposite. An SVM training algorithm could be a non-probabilistic, binary, linear classifier, although methods like Platt scaling exist to use SVM in a very probabilistic classification setting. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what's called the kernel trick,implicitly mapping their inputs into high-dimensional feature spaces. SVM also uses another method called Soft Margin which allows SVM to make certain number of mistakes and keep the margin as wide as possible so that other points can be classified correctly.

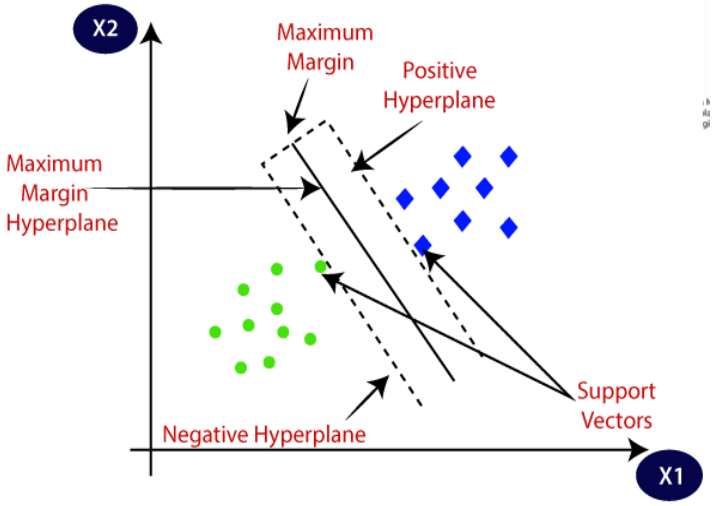

Fig 5: Support Vector Machine(SVM)

4.1.2.K-Nearest Neighbor(KNN)

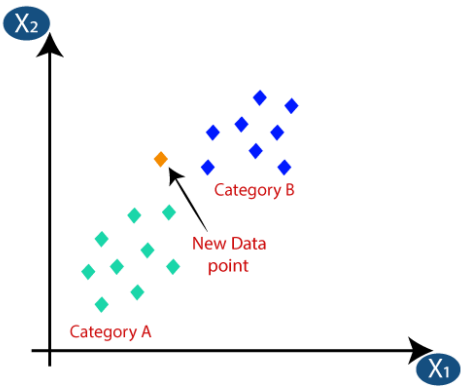

Fig:6 K-NN diagram

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

- K-NN algorithm stores all the available data and classifies a new data point based on the similarity.

- This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

4.1.3. LIGHT GBM

Light GBM is a gradient boosting framework which increases the potency of the model and reduces memory usage. Light GBM is a quick distributed, high performance framework based decision tree algorithmic program

Used for ranking, classification and many other machine learning task. The significance of light GBM are quicker training speed and higher potency, better accuracy than any other boosting algorithms, capable of large number of data, compatibility with enormous dataset.

Fig 4: Machine Learning Models

Model no	**Model**	**Module**	**Package**	**Tuning Parameters**
Model 1	Support vector classifier	SVC	sklearn	none
Model 2	K-NN	k-NN classifier	sklearn	none
Model 3	Light GBM	Light GBM classifier	Sklearn	none

Title of the paper	Author	Methology & Algorithm	Year	Merits	Demerits
An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks.	Faris,H.,Ala'M, A.Z.,Heidari, A.A., Aljarah, I., Mafarja,M., Hassonah,M.A. and Fujita,H., 2019.	Deep Learning	2019	Decision	Implement problem
Twitter spam detection:A systematic review	Abkenar, S.B., Kashani, M.H., Akbari, M. and Mahdipour, E., 2020	Machine Learning	2020	Flexible	Data train is hard
A deep learning model for Twitter spam detection.	Alom, Z., Carminati, B. and Ferrari, E., 2020.	Deep Learning	2020	Unthreatned	Less secure
Optimizing semantic LSTM for spam detection.	Jain, G., Sharma, M. and Agarwal, B., 2019.	LSTM	2019	Access data is easy	Access data is hard
A neural network-based ensemble approach for spam detection in Twitter.	Madisetty, S. and Desarkar, M.S., 2018	Neural network	2021	High prediction	Less prediction

5.Related works

6. RESULTS AND DISCUSSION

The proposed approach detects the spam parameters causing the IOT devices to be effected. To get the best results, the IOT dataset is used for the validation of proposed approach as described in the next Section. Transformations of Principal Components

6.1. Data Collection

We have collected the smart home dataset by REFIT project which is sponsored by Lough borough University. A total of twenty homes were used and advised to deploy the smart home technologies. The complete survey was conducted by the team of researchers. The experiments are varied from room to room, depending upon climate changes, floor plans,Internet supply and other attributes.The internal environmental conditions were captured using different sensors. There were more than 100,000 data points in each home for sensor monitoring. The survey was continued for almost 18 months. This dataset is openly available .

6.2. Experimental setup

To perform the experiments, we use the data set traces from the source as mentioned . Then, we performed
the experiments on RStudio (openly free software available). The software requirements are, Operating systemWindows 7/8/10 or MacOS 10.12+ or Ubuntu 14/16/18 or
Debian 8/10. Following are the results obtained.

6.3. Impact of data preprocessing

The preprocessing involves the selection of appliances being considered for the detection of spam parameters. The main idea is to find the various spam causing factors. Firstly, the feature reduction is done. The method used for feature reduction is the Principal Component Analysis (PCA), which reduces the dimensions of data. It results in series of Principal components (PC) which corresponds to each row with each column. In the IoT dataset used in this proposal, we have 15 features, so 15 PCs are generated as shown in Table VI. The pca() works in such a way that it reduces the variance among the features. The standard deviations of PCs is presented in and the transformations of PCs is presented.

7. CONCLUSION

The proposed framework, detects the spam parameters of IOT devices using machine learning models. The IOT dataset used for experiments, is pre-processed by using feature engineering procedure. By experimenting the framework with machine learning models, each IOT appliance is awarded with a spam score. This refines the conditions to be taken for successful working of IOT devices in a smart home. In future, we are planning to consider the climatic and surrounding features of IoT device to make them more secure and trustworthy.

REFERENCE

[1] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, “Iot security: ongoing challenges and research opportunities,” in 2014 IEEE 7th international conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.

[2] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, “Blockchain for iot security and privacy: The case study of a smart home,” in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

[3] E. Bertino and N. Islam, “Botnets and internet of things security,” Computer, no. 2, pp. 76–79, 2017.

[4] C. Zhang and R. Green, “Communication security in internet of thing: preventive measure and avoid dos attack over IOT network,” in Proceedings of the 18th Symposium on Communications &

Networking. Society for Computer Simulation International, 2015, pp. 8–15.

[5] W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the internet: Attacks, costs and responses," Information systems, vol. 36, no. 3, pp. 675–705, 2011.

[6] H. Eun, H. Lee, and H. Oh, "Conditional privacy preserving security protocol for nfc applications," IEEE Transactions on Consumer Electronics, vol. 59, no. 1, pp. 153–160, 2013.

[7] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in 2009 International Joint Conference on Neural Networks. IEEE, 2009, pp. 1680–1687.

[8] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996– 2018, 2014.

[9] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2015.

[10] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," Soft Computing, vol. 20, no. 1, pp. 343–357, 2016.

[11] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, "A system for denial-of-service attack detection based on multivariate correlation analysis," IEEE transactions on parallel and distributed systems, vol. 25, no. 2, pp. 447–456, 2013. [12] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "Sinr-based dos attack on remote state estimation: A game-theoretic approach," IEEE Transactions on Control of Network Systems, vol. 4, no. 3, pp. 632–642, 2016.

[13] L. Xiao, Y. Li, X. Huang, and X. Du, "Cloud-based malware detection game for mobile devices with offloading," IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2742–2750, 2017.

[14] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," Knowledge and information systems, vol. 34, no. 1, pp. 23–54, 2013.

[15] I. Jolliffe, Principal component analysis. Springer, 2011.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, pp. 1157–1182, 2003.

[17] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.

[18] A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial intelligence driven mechanism for edge computing based industrial applications," IEEE Transactions on Industrial Informatics, 2019.

[19] A. H. Sodhro, Z. Luo, G. H. Sodhro, M. Muzamal, J. J. Rodrigues, and V. H. C. de Albuquerque, "Artificial intelligence based qos optimization for multimedia communication in iov systems," Future Generation Computer Systems, vol. 95, pp. 667–680, 2019.

[20] L. University, "Refit smart home dataset," https://repository.lboro.ac.uk/ articles/REFIT Smart Home dataset/2070091, 2019 (accessed April 26, 2019).

[21] R, "Rstudio," 2019 (accessed October 23, 2019)