JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Intrusion Detection System based on optimal feature selection and classifiers

[1]Ikshita Bansal, [2]Rashim Rana
[1,2]Department of CSE., HIET, Kala-amb, Himachal Pardesh, India

Abstract: In today's environment the demand of users for accessing and storing data at high speed may cause the security of data. There are many intruders those violates the security of system by spreading the various kinds of attacks. Although various researchers proposed different ID mechanism by applying different classification algorithms but the anomaly detection rate is less in some algorithms and accuracy is also low in some algorithms. To overcome these kinds of problems in this paper an attempt has been made to proposed to Collaborative Filtering based mechanism to detect anomalies with high detection rate.

IndexTerms **- IDS; Machine learning; Random Forest and attacks.**

I. INTRODUCTION

Today, there is a difficult issue for PC researchers and professionals for detection and avoidance attacks and it have become a significant focal point of as PC attacks have become an expanding danger to business just as our everyday lives[1]. Intrusion exercises are expanding because of the expansion of organization use. Lately, attacks on PC systems are expanding and require viable and proficient intrusion detection system. Objective of IDS is to discover intrusion in ordinary review data[2].
Intrusion detection system is plan to screen the occasions in a system or organization by determining if is an intrusion. It additionally screen the organization traffic for dubious movement and alarm the organization or system chairman about those attacks when happened[3]. The target of this system expect to cover the accessibility, secrecy and uprightness of basic organized data system.
Specialists have created two principle approaches for intrusion detection: abuse and inconsistency intrusion detection. Abuse comprises of addressing the particular examples of intrusions that adventure known system weaknesses or disregard system security approaches [4].
On the opposite side, irregularity detection expects that all meddlesome exercises are essentially strange. This implies that in the event that we could set up a typical movement profile for a system, we could, in principle, banner all system states fluctuating from the set up profile as intrusion endeavours. These two sorts of systems have their own qualities and short comings[5].
The previous can identify known attacks with an exceptionally high exactness by means of example coordinating on known marks, yet can't distinguish novel attacks on the grounds that their marks are not yet accessible for design coordinating[6]. The last can identify novel attacks however overall for most such existing systems, have a high bogus alert rate since it is hard to produce viable typical conduct profiles for secured systems. Along these lines, we propose intrusion detection system utilizing Random forest[7-8].

II MACHINE LEARNING

Machine learning shows up as the gadget needed for deal with or deal with the enormous measure of data created in IT. It fits in as the last piece of the IT system which is driven by data assortment, examination, and dynamic. Machine learning strategies give a proficient method to dissect, and afterward settle on suitable choices to identify intrusions[9-10].
Machine learning (ML) is a term which alludes to learning and making forecasts from accessible data by a system. It is involved different calculations which dissect the accessible data through a bunch of guidelines to create data-driven expectations or potentially choices. Machine learning goes through the thorough interaction of planning and programming unequivocal calculations with anticipated execution[11-13].

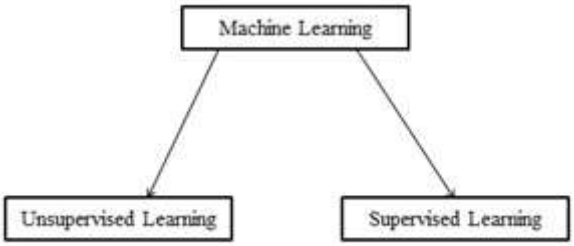

Fig.1. Types of Machine Learning[13]

Supervised Learning: Supervised learning is where the input variables and the output variables are already determined, and a learning algorithm is used to learn how to map the input to the output[14].
Unsupervised Learning: In this learning, the input data does not contain any information of corresponding output variables[14].

III PROPOSED WORK

In this paper, a collaborative filtering approach is proposed. In the proposed approach KDD dataset is used. The dataset is downloaded from Kaggle.com. Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users[15-16].

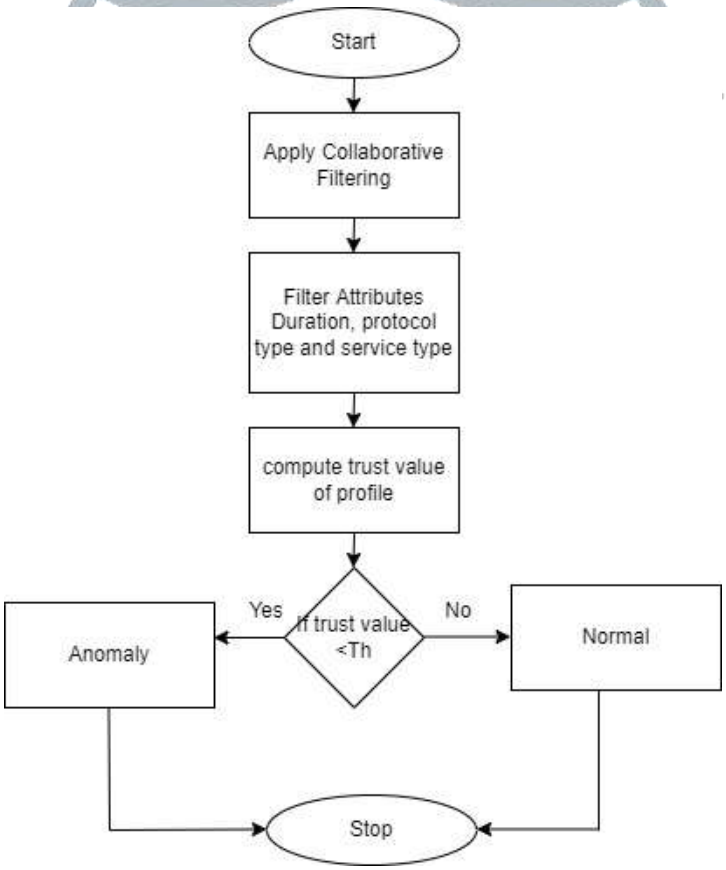

Fig.2. Flow Diagram of Proposed Mechanism

In proposed mechanism, take the average rating given by a user, an average rating given to a particular instance of attack and overall average of ratings for the entire instances of attack[17].

In Prediction using Correlation based on Similarity, Collaborative Filtering Based Approach is used. Here it is assumed that user (u) to whom recommendations are to be made, find a group of other users whose behaviour are similar to the user (u) and call these set of users, Neighbourhood of user(u). Then, find instances of attacks and distinguish into classes normal and anomalies[18-20].

Fig.2 illustrates the working of the proposed mechanism. Here first collected real dataset is faded into the database and distinguish arguments into two categories called the argument for training and argument for testing. In testing phase arguments are used to test the profiles of attack. In the training, phase parses the dataset and calculate correlated matrix, find mean and create a prediction of an instance based on correlation matrix and mean parameters. After that apply collaborative filtering to distinguish attributes such as duration, protocol type and service type to calculate trust value of each profile. If trust value is grater, then some threshold value then it is normal profile otherwise it is anomaly profile. The anomaly profiles are detected based on past history of profiles in this way proposed mechanism can distinguish normal and anomaly profiles.[20].

IV RESULTS AND ANALYSIS

4.1 Tool used: We configure Weka in eclipse to implement proposed mechanism. To analyse proposed mechanism KDD dataset is used. This dataset is open-source dataset and contains various attributes such as service type duration and protocol type.

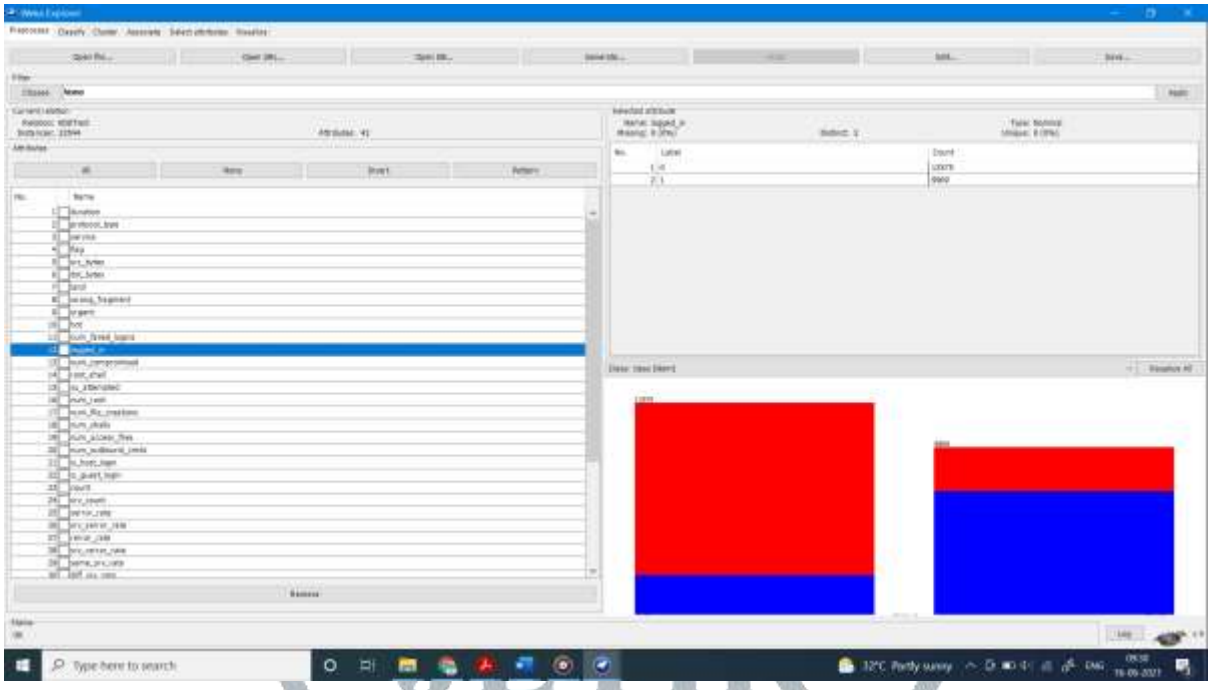

Fig.3. Classes used in Dataset

4.2 Performance Metrices

- *Accuracy*: It depicts the total number of positive (+ve) outcomes in perspective of the total number of negative (-ve) outcomes from complete dataset. The formula for calculation is as follows:

$$Accuracy = TP+TN/ (TP+FP+FN+TN)$$

- *Confusion Matrix*: It depicts the prediction values of data in terms of TP, TN, FN, FP i.e., true +ve, true –ve, false +ve and false -ve. On the basis of these parameters the sensitivity and accuracy of techniques has been computed.
- *F-Measure*: It depicts the combination of precision and recall values. If F-score is high then accuracy of classification is high.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

<table>
<tr><th>Techniques</th><th>Confusion matrix</th></tr>
<tr><td>Logistic Regression</td><td>
<table>
<tr><td>a(Functional)</td><td>b(Nonfunctional)</td><td>Class Tested</td></tr>
<tr><td>9078</td><td>633</td><td>Normal</td></tr>
<tr><td>393</td><td>12440</td><td>Anomaly</td></tr>
</table>
</td></tr>
<tr><td>Naïve Bayes</td><td>
<table>
<tr><td>a(Functional)</td><td>b(Nonfunctional)</td><td>Class</td></tr>
<tr><td>9225</td><td>486</td><td>Normal</td></tr>
<tr><td>358</td><td>8975</td><td>Anomaly</td></tr>
</table>
</td></tr>
<tr><td>Random Forest</td><td>
<table>
<tr><td>a(Functional)</td><td>b(Nonfunctional)</td><td>Class</td></tr>
<tr><td>9564</td><td>147</td><td>Normal</td></tr>
<tr><td>164</td><td>12669</td><td>Anomaly</td></tr>
</table>
</td></tr>
<tr><td>Proposed</td><td>
<table>
<tr><td>a(Functional)</td><td>b(Nonfunctional)</td><td>Class</td></tr>
<tr><td>9770</td><td>133</td><td>Normal</td></tr>
</table>
</td></tr>
</table>

	172	13778	Anomaly

Table 1. Confusion matrix Comparison of Proposed, Random Forest, Naïve Bayes and Logistic Regression

Table1 shows confusion matrix comparison of all standard recommendation techniques in perspective of proposed technique. The confusion matrix is computed based on TP FP rates of computation. In propose mechanism the accurate classification of anomaly is high as compare to other techniques.

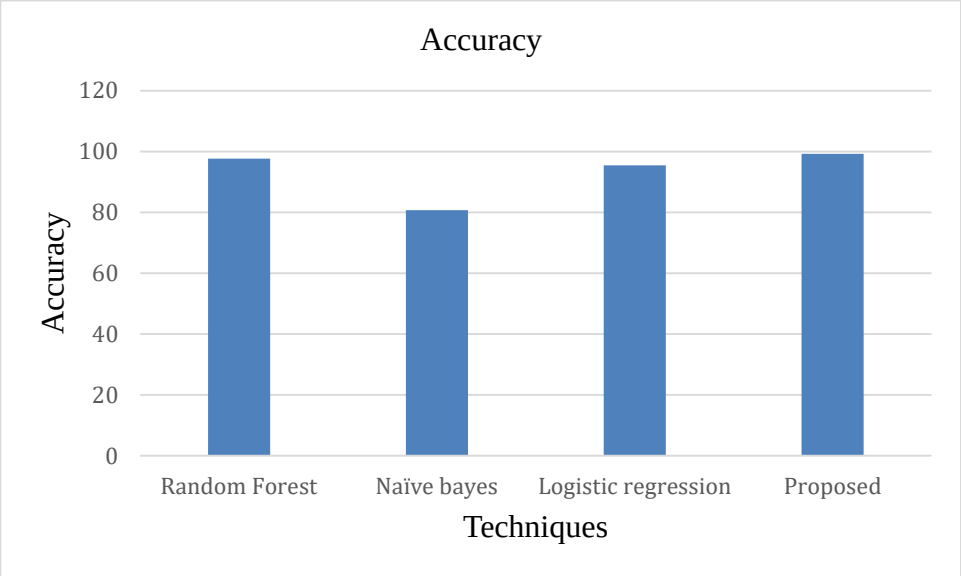

Fig.4. F-Accuracy comparison

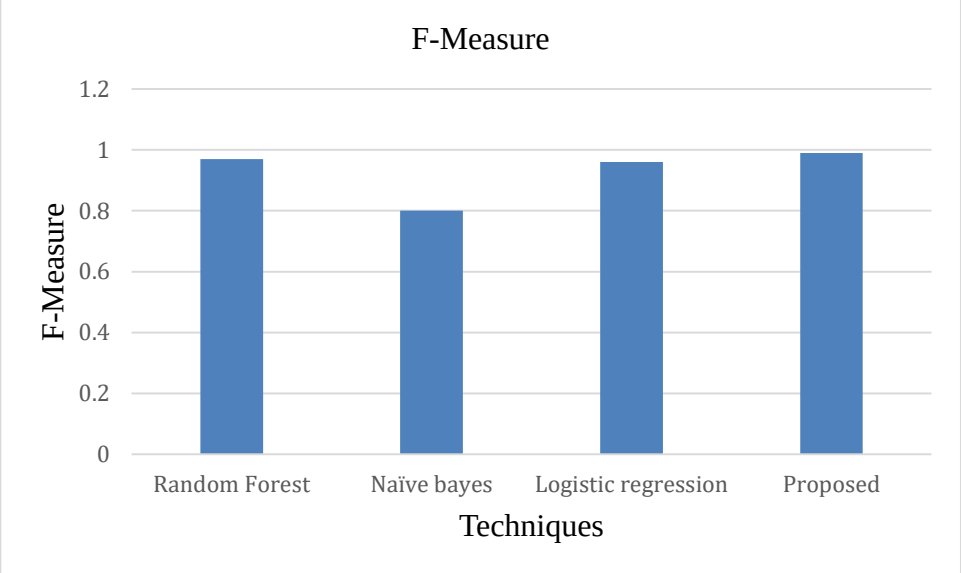

Fig .5. Measure score comparison

Fig .4 depicts accuracy comparison of proposed techniques in perspective of other techniques such as naïve bayes logistic regression and random forest. In proposed mechanism the rate of accurate anomaly detection is high as comparison to other techniques also in naïve bayes technique the accuracy is very less among all techniques whereas Fig.5 depicts comparison of proposed technique with other existing techniques in perspective of F-measure. In proposed mechanism F-measure value is high in comparison to other techniques whereas naïve bayes have low F-measure score in comparison to all other techniques.

V CONCLUSION

In this paper, a collaborative filtering approach is proposed. In the proposed approach KDD dataset is used. The dataset is downloaded from Kaggle.com. Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. In proposed mechanism, take the average rating given by a user, an average rating given to a particular instance of attack and overall average of ratings for the entire instances of attack.In proposed mechanism the rate of accurate anomaly detection is high as comparison to other techniques also in naïve bayes technique the accuracy is very less among all techniques whereas in proposed mechanism F-measure value is high in comparison to other techniques whereas naïve bayes have low F-measure score in comparison to all other techniques.

REFERENCES

[1] Anderson, P. 1980. Computer Security Threat Monitoring and Surveillance, Technical Report, James Anderson Report, Pennsylvania.

[2] Nadiammai, G. V., Krishnaveni, S. and Hemalatha, M. 2011. A Comprehensive Analysis and Study in IDS Using Data Mining Techniques. IJCA, 35: 51–56.

[3] Breiman, L. 2001. Random Forests, Machine Learning, 45(1), 5–32.

[4] Buczak, A. L. and Guven, E. 2016. A survey of data mining andmachine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18: 1153–1176.

[5] Kolosnjaji, B., Eraisha, G., Webster, G., Zarras, A. and Eckert C. 2017. Empowering convolutional networks for malware classificationand analysis. in Proceedings of the International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA: pp. 3838–3845.

[6] Li, D., Chen, D., Goh, J. and Ng, S. 2019. Anomaly detection withgenerative adversarial networks for multivariate time series.

[7] Chen, J. Sathe, S. Aggarwal, C. and Turaga, D. 2017. Outlier detection with autoencoder ensembles. in Proceedings of SIAM International Conference on Data Mining, Houston, TX, USA: 90–98.

[8] Erfani, S. M. Rajasegarar, S. Karunasekera, S. and Leckie, C. 2016. High-dimensional and large-scale anomaly detection using alinear one-class SVM with deep learning. Pattern Recognition 58: 121–134

[9] Khalvati, L. Keshtgary, M. and Rikhtegar, N. 2018. Intrusion detectionbased on a novel hybrid learning approach. Journal of AI and Data Mining 6(1): 157–162.

[10] Vinayakumar, R. Soman, K. P. and Poornachandran P. 2017. Applying convolutional neural network for network intrusiondetection. in Proceedings of the International Conference on Advances in Computing Communications and Informatics(ICACCI), Udupi, India: 1222–1228.

[11] Potluri S. and Diedrich C. 2016. Accelerated deep neural networksfor enhanced Intrusion Detection System. in Proceedings of IEEE 21st International Conference on Emerging Technologya nd Factory Automation (ETFA), Berlin,Germany: 1–8.

[12] Gao, N., Gao, L., Gao, Q., Nabila, HFarnaaz, and Jabbar. M. A. 2016. Random forest modeling for network intrusion detection system. Procedia Computer Science 89: 213-217.

[13] Mahsa, Ebrahimian, and RashaKashef 2020. Efficient Detection of Shilling's Attacks in Collaborative Filtering Recommendation Systems Using Deep Learning Models. In Proceeding of IEEE International Conference on Industrial Engineering and Engineering Management (IEEM): 460-464.

[14] Sundar, Agnideven Palanisamy, Li, Feng, Zou, Xukai, Gao, Tianchong, and Russomanno, Evan D. 2020. Unnderstanding shilling attacks and their detection traits: a comprehensive survey. IEEE Access 8: 171703-171715.

[15] Rezaimehr, Fatemeh, and Chitra, Dadkhah 2021. A survey of attack detection approaches in collaborative filtering recommender systems. Artificial Intelligence Review 54(3): 2011-2066.

[16] Panagiotakis, Costas, Harris Papadakis, and ParaskeviFragopoulou 2020. Unsupervised and supervised methods for the detection of hurriedly created profiles in recommender systems. International Journal of Machine Learning and Cybernetics 11(9): 2165-2179.

[17] Alabdulrahman, Rabaa, and Herna Viktor 2021. Catering for unique tastes: Targeting grey-sheep users recommender systems through one-class machine learning. Expert Systems with Applications 166: 1-12.

[18] Zafar Ali Khan, N., and R. Mahalakshmi 2021. Hybrid Collaborative Fusion Based Product Recommendation Exploiting Sentiments from Implicit and Explicit Reviews. Journal of Interconnection Networks: 2141013.

[19] Yin, Chunyong, Lingfeng Shi, Ruxia Sun, and Jin Wang 2020. Improved collaborative filtering recommendation algorithm based on differential privacy protection. The Journal of Supercomputing 76(7): 5161-5174.

[20] Yang, Zhihai, Qindong Sun, Yaling Zhang, and Beibei Zhang 2018. Uncovering anomalous rating behaviors for rating systems. Neurocomputing 308: 205-226.

[21] Dan, Wu. 2021. Intelligent English resource recommendation and teaching effect based on symmetric SDAE collaborative filtering algorithm. Journal of Ambient Intelligence and Humanized Computing: 1-11.

AUTHORS PROFILE

Ikshita Bansal is pursuing Master of Technology (Computer Science & Engineering) from Himalayan institute of Engineering & Technology, Kala-amb, H.P., India. She has completed her Bachelor of Technology (Computer Science & Engineering) from Eternal University, Baru Sahib, H.P., India in 2012.

Rashim Rana, M.tech., is an Assistant Professor at Himalayan institute of Engineering & Technology, Kala-amb, H.P., India. She has good research background. She is very hardworking and co-operative.