Research Journal of Computer Systems and Engineering

ISSN: 2230-8571, 2230-8563

Volume 03 Issue 01 - 2022 (January to June)

Page 62:66

Social Network Based Privacy Data Optimization Using Ensemble Deep Learning Architectures

Dr. N. Shanthi
Life Science and Mathematics
Pachaiyappa's College, Affiliated to University of Madras
https://orcid.org/0000-0002-1261-2499
prithishanthi@gmail.com

Shreyas J
Computer Science & Communication Engineering
Crrencia Technologies Pvt. Ltd.
ORCHID ID-0000-0002-3427-4486
shreyasj.email@gmail.com

Article History

Received: 22 January 2022
Revised: 14 April 2022
Accepted: 19 May 2022

Abstract

As a result of changes in technology, enormous amounts of data are produced every second. A tremendous amount of data is generated every second by social networking and data mining. This paper proposed the novel intrusion detection based data optimization by deep learning techniques. Initially the security has been improved by using Ensemble algorithm combining decision tree with clustering based IDS. The network generates data at a high rate, volume, and variety, making it highly challenging to identify assaults using conventional methods. So the data has been optimized using KNN classification which improves the data accuracy. We developed a neural network for detecting malicious user attacks using a deep-learning technique. We discovered that a deep learning model could improve accuracy such that a social network's ability to mitigate attacks is as effective as possible.

Keywords: Social networking, data mining, data optimization, deep learning techniques, clustering, classification.

1. Introduction:

Web-based services that let users build public or semi-public profiles within a domain so they can interact with other users within the network are referred to as social networks [1]. The creation and exchange of User-Generated Content via social networks has enhanced the idea and technology of Web 2.0 [2]. It has been discovered that data mining approaches can handle the three main disagreements with social network data, namely scale, noise, and dynamism. Because social network datasets are so large, automated information processing is necessary to analyse them quickly [3]. But DL talks about supervised or unsupervised machine learning methods that automatically pick up the hierarchical categorization representations. Due to its inspiration from scientific findings of how the human brain processes information, DL has recently attracted considerable attention from the research community [4].

2. Related works:

Many different tasks can benefit from the use of machine learning techniques. This research [5] mainly retrieved four significant variables, namely the tweets, time of publication, language, geo-position, and Twitter client, after gathering data from various sample Twitter profiles. The performance of the current classification techniques, including Random Forest [7], J48 [8], CNN [9], and Sequential Minimal Optimization, is compared in Paper [6]. (SMO). [10] presents research on the detection of phoney profiles using a variety of data mining approaches on the LinkedIn dataset.

3. Research methodology:

The suggested feature selection model attempts to improve social network performance. Many academics have used data mining and machine learning approaches in recent years to solve issues and improve system performance. In order to increase the effectiveness of social networks, this work used the ensemble technique to identify malevolent individuals. The architecture of the suggested model is shown in Figure 3. The steps of the proposed model are specifically identified in the following sections.

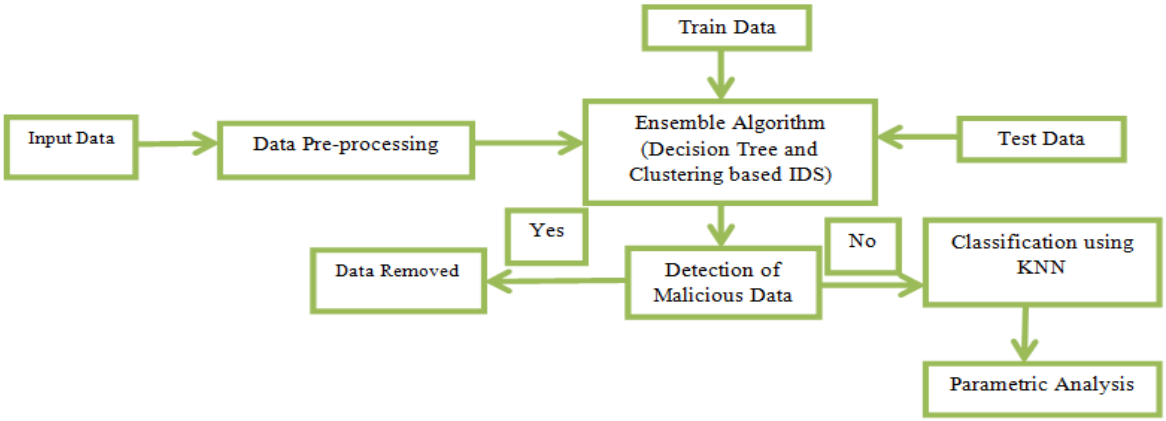

Figure1: Overall proposed architecture

The conversion of a URL into a feature vector, where many types of information can be taken into consideration and various methodologies can be applied, is the first crucial step. This part cannot be simply computed by a mathematical function, unlike learning the prediction model (not for most of it). By crawling all pertinent information about the URL, a feature representation is built using domain knowledge and associated skills.

3.1 Decision tree algorithm:

The CART (Classification and Regression Trees) is a decision tree classifier applied to determine the classified output with high degree of accuracy.

3.2 Tree-growing algorithm:

BEGIN: Assign all training data to the root node

Define the root node as a terminal node

SPLIT:

New_splits=0

FOR every terminal node in the tree:

If the terminal node sample size is too small or all instances in the node belong to the same target class goto GETNEXT

New_splits+1

GETNEXT:

NEXT

3.3 Pruning algorithm:

DEFINE: r(t)= training data misclassification rate in node

t p(t)= fraction of the training data in node

*t R(t)= r(t)*p(t)*

t_left=left child of node t

t_right=right child of node t

|T| = number of terminal nodes in tree T

BEGIN:Tmax=largest tree grown

Current_Tree=Tmax

For all parents t of two terminal nodes

Remove all splits for which R(t)=R(t_left) + R(t_right)

Current_tree=Tmax after pruning

3.4 Decision tree algorithm:

Input Training set $\{(x_1, y_1, \ldots., (x_n, y_n)\}$

i. Initialize T to be a single unlabeled node.

ii. **While** there are unlabeled users in *T* **do**

iii. Navigate data samples to their corresponding users.

iv. **for all** unlabeled users*v* in *T* **do**

v. **if** *v* satisfies the stopping criterion **or** there are no samples reaching *v* **then**

vi. Label *v* with the most frequent label among the samples reaching *v*

vii. **else**

viii. Choose cluster splits for *v* and estimate D for each of them.

ix.**end if**

x.**end for**

xi.**end while**

The output of decision tree will show the whether the cluster iterated is malicious or not. When the user in a cluster is detected to be malicious will be removed from the cluster and the data will be reported to be malicious URL and it will be removed. When the user is detected to be normal user, then the data of the URL has been classified using KNN classification.

4. Performance Analysis:

The ensemble classifier with clustering based IDS was trained and classified using Python and the Natural Language Tool Kit. We used a data set of 19340 bytes, of which 1000 bytes were used for testing and 18340 bytes for training. Twitter API is used to automatically gather tweets, which are then manually classified as good or negative. 600 uplifting and 600 depressing tweets are combined to form a dataset.

4.1 Parametric comparison of proposed with existing technique:

This section discuss about the parametric comparison between existing and proposed technique. The results are shows below. The parameter compared is accuracy, precision, recall and F-1 score. On comparing with exiting technique, proposed DEC_TREE WITH IDS-KNN enhanced output in terms of accuracy, precision, recall and F-1 score. The below table-5 shows the overall parametric comparison of proposed DEC_TREE WITH IDS-KNN and figure 10 shows the graphical representation. By this overall comparison graph and table, it is proved that the proposed technique identify the malicious user with higher accuracy.

Table-5 Overall Parametric comparison of Proposed DEC_TREE WITH IDS-KNN

Parameters	PRO_DEC_TREE WITH IDS-KNN
Accuracy	90.2
Precision	85.8
Recall	80.2
F1-Score	70.2

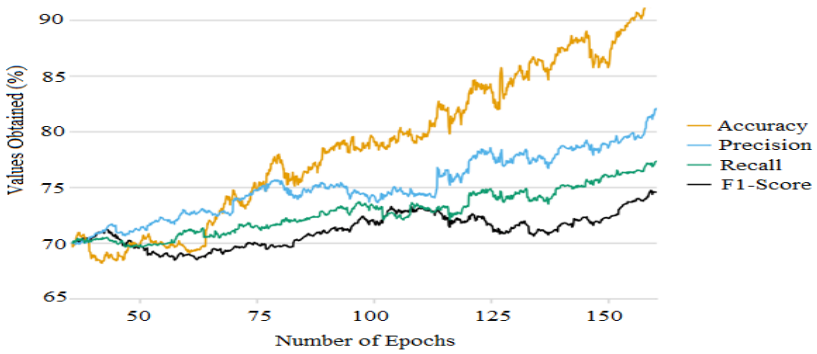

Figure- 10 Overall Parametric comparison of Proposed DEC_TREE WITH IDS-KNN

5 Conclusion:

The number of gadgets connected to the internet is rapidly rising as the internet has ingrained itself into every aspect of modern life. However, some issues are getting worse, and it's unclear how to

solve them. So this proposed technique gives the secured data transmission with optimization. Here the clustering based IDS with decision tree algorithm has been used as ensemble technique. This ensemble output will detect the malicious user of the social network by the decision tree algorithm. When the malicious user is detected then the data will be removed before the classification process and normal user will be sent for classification in achieving higher data accuracy. KNN classifier has been used and this has enhanced the classification accuracy as shown in experimental results. In future this work can be extended by increasing the data capacity and can be implemented various classifier of neural network.

Reference:

[1] Islam, MdRafiqul, et al. "Depression detection from social network data using machine learning techniques." *Health information science and systems* 6.1 (2018): 1-12.

[2] Kim, Jooho, and MakarandHastak. "Social network analysis: Characteristics of online social networks after a disaster." *International Journal of Information Management* 38.1 (2018): 86-96.

[3] B. T. Pham, D. Tien Bui, I. Prakash, L. H. Nguyen, and M. B. Dholakia, "A comparative study of sequential minimal optimizationbased support vector machines, vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS," Environ. Earth Sci., vol. 76, no. 10, 2017

[4] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," Comput. Electr. Eng., vol. 65, no. 3, pp. 165–177, 2018.

[5] H. Zheng and C. Wu, "Predicting personality using facebook status based on semi-supervised learning," ACM Int. Conf. Proceeding Ser., vol. Part F1481, pp. 59–64, 2019.

[6] M. Gaikar, J. Chavan, K. Indore, and R. Shedge, "Depression Detection and Prevention System by Analysing Tweets," SSRN Electron. J., pp. 1–6, 2019.

[7] Xu, Xiaodong, et al. "Incorporating machine learning with building network analysis to predict multi-building energy use." *Energy and Buildings* 186 (2019): 80-97.

[8] Maione, Camila, Donald R. Nelson, and Rommel Melgaço Barbosa. "Research on social data by means of cluster analysis." *Applied Computing and Informatics* 15.2 (2019): 153-162.

[9] Faker, Osama, and ErdoganDogdu. "Intrusion detection using big data and deep learning techniques." *Proceedings of the 2019 ACM Southeast Conference*. 2019.

[10] Sultana, Nasrin, et al. "Survey on SDN based network intrusion detection system using machine learning approaches." *Peer-to-Peer Networking and Applications* 12.2 (2019): 493-501.