Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.778065

STRUCTURAL PATTERNS CLASSIFICATION AND PREDICTION IN DARK WEB MINING USING RECURRENT NEURAL NETWORK

G.Dhivya, Dr.M.Mohankumar

[1]Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, (Deemed to be University, Coimbatore, India – 641021)

[2]Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education, (Deemed to be University, Coimbatore, India – 641021)

E-mail:-dhivya.gurusamy@kahedu.edu.in, mohankumarcs@kahedu.edu.in

Abstract

Dark Web patterns in structural patterns mining increases a number of issues (including a lot of unessential and avoidable information), which in turn boosts a many cybercrimes like illegal trade, child abusement, criminal activity, and unlawful online purchasing. Structural patterns of this study is the danger level of dark web mining is predicted using the Naive Bayes and Decision Tree algorithms. Using datasets from the ToR, a system forecasts the patterns. Because of the more variety of data, it might be difficult to analyse these illegal data in online. In order to investigate a current request for improving the structured data as a client profile, a technique for evaluating criminal conduct is required. Structural Patterns mining in the Dark Web forum contains multi-dimensional data sets that yield suspicious findings. Uncertain categorization outcomes are the root of this forum.

Keywords: Dark Web, ToR, RNN, SVM, Cyber Attacks

Problem Definition

In today world, the enormous attacks in Dark Web Forum. Cyber Security contains large information, which is difficult to process by manual methods, Data Science is the challenging one because data contains large amount of records. It is proposed to develop centralized cyber attacks in Dark Web system using Data Science. In the cyber security large set of records take as input and it is aimed to extract the needed information from the record by using Machine Learning Algorithms. It is very difficult to find Cyber attacks of Dark Web Forum and its structure.

Introduction

The expert and experienced Scientists are not also available against the large population sometimes cyber datasets are being neglected. The existing system is not able to extract all the information and knowledge from datasets. Complex query for the Dark Web Structure Identification is very difficult to analyze the Cyber Attacks in Dark Web Forum. The proposed system is named as “Patterns Classification and Prediction in Dark Web Mining”. In this system, overcomes all these limitation and hiding patterns of the existing system called as unlabelled data will be changed to Labeled data by using RNN based Machine Learning Algorithms.

Logistic Regression Algorithms

Logistic regression is a statistical method, which is useful for Machine Learning. For binary classification issues (i.e., values of two class with issues), Logistic regression is the quality approach. The goal of both linear and non-linear logistic regression determines coefficient values that weigh every input variable. The logistic function is a non-linear function that modifies predictions to produce distinct results. The logistic role resembles a value S which will reconstruct the range from 0 to 1. The result of the logistic function will be used to forecast the class value and a snap value to 0 and 1.

Logistic Regression

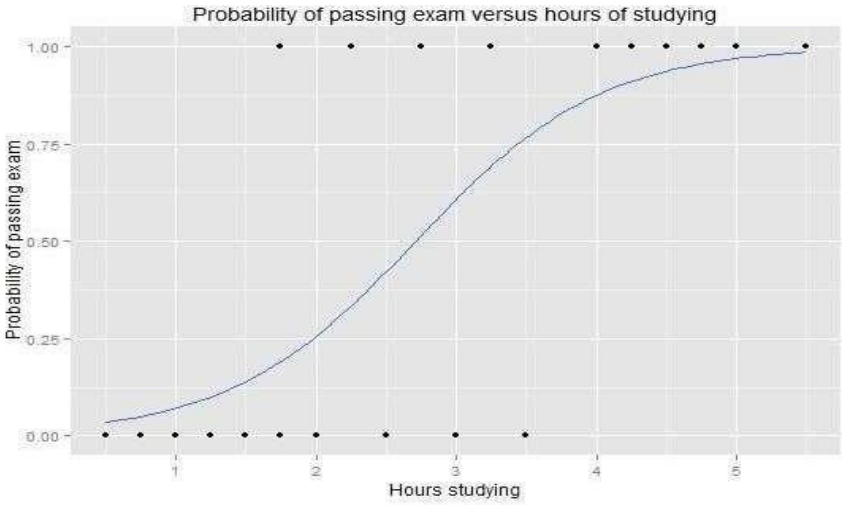

Figure 1 Probability Chart

Recurrent Neural Network

The RNN algorithm is very easy to use and quite powerful. The full training dataset serves to the model representation for RNN. The entire training set will be searched for R instances which are neighbors. While comparing R instances—the predicted variable for those R instances is summed in order to create predictions for the incoming data point. This could be the modal (or most often) class value in a classification task, or the mean which is called as output variable of the regression problem.

Understanding the degree of similarity between the data instances is the difficulty. In cases where all the attributes are on the same measure like inches. For example, the simplest method is the Euclidean distance, which can be computed immediately from the input variable differences.

R-Nearest Neighbors

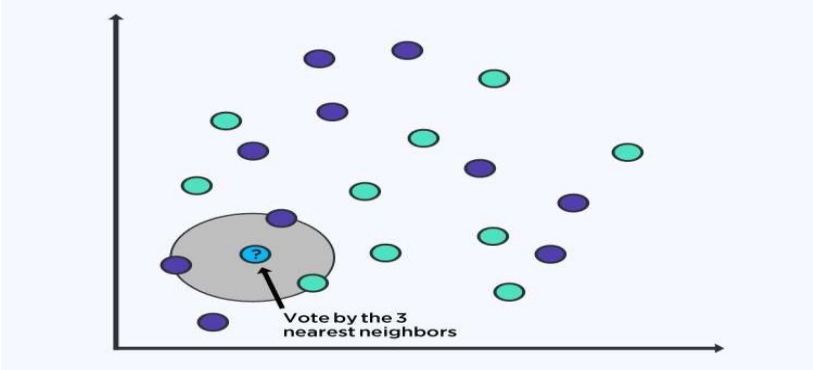

Figure 2 R-Nearest Neighbors

Support Vector Machines Algorithm

Backing the vector machine is the most well known as well as extensively discussed machine learning techniques. The input variable space splitup is called a hyperlane. To classify the points in the input variable space into classes 0 or 1, the best hyperlane is chosen by using Support Vector Machine. Assume for the moment that a line representing this in two dimensions can completely separate each of our input points. The SVM learning technique finds the coefficient that offers the best hyperlane partition of the classes.

Support Vector Machines

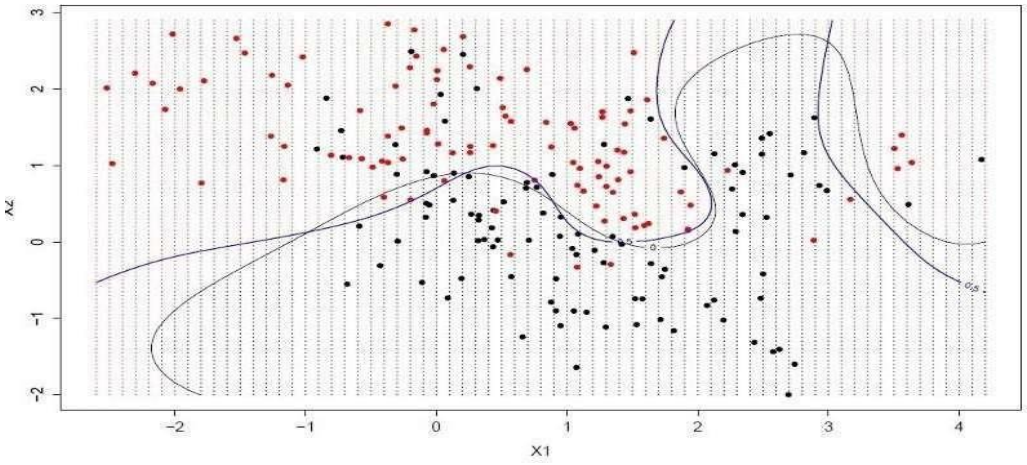

Figure 3 Support Vector Machines

Naïve Bayes Algorithm

One of the most predictive modeling technique is called Naïve Bayes algorithm. By using training data this model is classified into two probabilities:

They are 1) Every Class probability; and 2) Condition x value of probability in each class. Once established, the Bayes Theorem applies the probability model to generate predictions for fresh data. In real-world data, the common practice is the Gausssian distribution assumption. It may

a form of bell curve in order to facilitate fast estimation of probabilities. This algorithm is used to get unique value of each input variable. Though it is highly accepted the obstructive of actual facts, the approach is remarkably effective for a wide range of complex problems.

.Naive Bayes

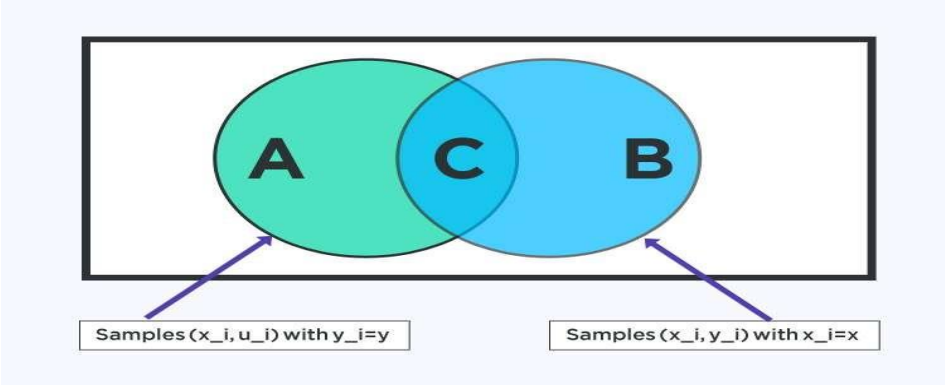

Figure 4 Naive Bayes Samples

Decision Tree Algorithm

The decision tree algorithm is a member of the family of supervised machine learning algorithms. It is applicable to problems involving both classification and regression. The main aim is to develop a model that forecasts the value of a target variable. This is accomplished by using a decision tree, in which the internal node of the tree represents attributes such as labelled or unlabeled data, and the leaf node corresponds to a class label.

Structure Of Decision Tree

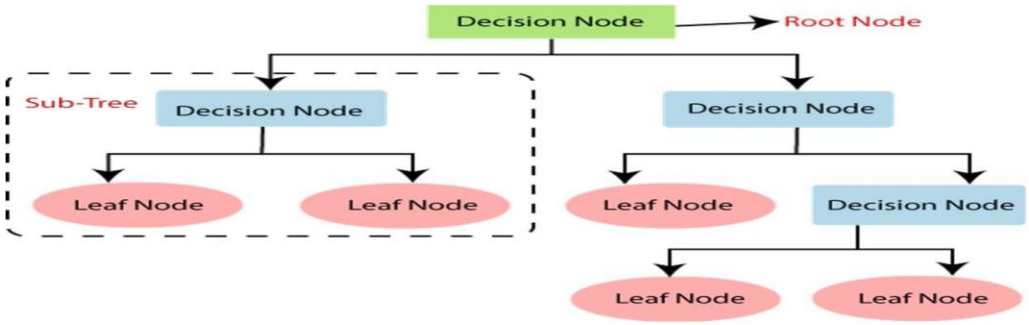

Figure 5 Nodes Description about Decision Tree

Random Forest Algorithm

Another one effective and successful machine learning algorithm is called Random Forest Algorithm. This kind of ensemble machine learning algorithm is known as "bagging," or Bootstrap Aggregation. An effective statistical technique for estimating a quantity from a sample of data is the bootstrap. Multiple samples of the data are gathered, and used to calculate the mean, mean values and average to have a better idea of the true mean value.

The same method is applied in bagging, but for estimating complete statistical models, most frequently decision trees. The training data is divided into many samples, and models are then built for each sample. Each model creates a prediction when you need to make one for new data.

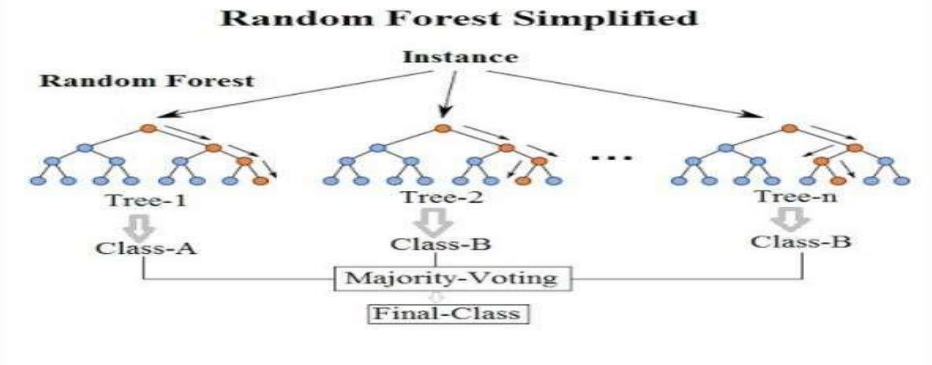

Figure 6 Random Forest Simplified

Because of this, the models developed each sample of the data are dissimilar with the previous. If not, they are nevertheless accurate in their own unique and special ways. The true underlying output value is more accurately estimated when the estimates are combined. A high variance algorithm (such as decision trees) that produces good results can frequently be bagged for even better outcomes.

Dataset Description (Cityscapes, Dark Zurich)

The proposed system uses CSV files for input (Comma-Separated Values). The CSV file includes Cityscapes, Dark Zurich are used to predict crawler information based on what we require. This is the year wise data detected from 2014 onwards. This graph represents how the RNN is used to detect the dark net crawler in Dark Web. The RNN Model is used to show crawling site of the users while comparing other models like LSTM –Long Short Term Memory is a type of Recurrent Neural Network.

Figure 7 Dataset Description

Methodology

- Predict whether a pattern is labeled or unlabelled data. If it is Positive (+) = 1, labeled data or Negative (-) = 0, unlabelled data.
- Try out various Classification Models to determine which produces the most accurate results.
- Analyze correlations and trends within the dataset.
- Identify the characteristics that Positive/Negative labeled data values the most.
- Comparative analysis of RNN with Logistic Regression, Decision Tree, Naïve Bayes and Support Vector Machine.

Correlation Matrix

Within seconds, the correlation matric can be determined whether a relationship exists between a variable and our predictor (target).Labeled and target (our predictor) have a positive connection. This makes sense because there is a higher likelihood of identifying patterns when there is more structural behavior. Text is Value 1, text size is Value 2, identical words are Value 3, and semantic analysis is Value 4.

Furthermore, observation of a negative connection between structured and unstructured data. When the unstructured data will communicate with the client the attacks might be happened. This unstructured data may result if the data arrival in the path compared to the semantic analysis.

Filtering data by positive & negative labeled data

It is clear that the means of several Darkweb Datasets differ significantly when positive and negative data are compared. We can see from looking at the specifics that the positive and negative numbers average. Additionally, positive individuals show roughly a third less ST depression during exercise compared to rest.

Prepare Data for Modeling

Just keep in mind the acronym ASN (Assign, Split, Normalize) when preparing data for modeling. Divide the set of data into the Training and Test set by assigning the other features to column X, and our classification predictor in the last column. Normalize: The distribution will be altered to have a mean of 0 and a standard deviation of 1 by normalizing the data.

Modeling/Training

On the training set, we will now train multiple classification models to see which produces the maximum accuracy. We will contrast the precision of Naives Bayes Classifier, Decision Trees, Random Forest, Support Vector Machine, and Logistic Regression. Finally, it may be concluded from a comparison of the six models that Model 6: RNN produces the highest accuracy with about 80% accuracy.

The F1-score is the harmonic mean of the precision and recall values. The F1 score can range from a minimum of 0 to a maximum of 1. Two times ((precision x recall) / (precision + recall)) is the F1 Score. The samples in the real response that classifies to that class as support.

Confusion Matrix

According to our data, there are 21 True Positives and 28 True Negatives.

There are three and nine errors total. Nine Type 1 errors, also known as False Positives, were found to have erroneously predicted a positive result. Three Type 2 errors (false negatives) exist, all of which you correctly predicted to be false.

As a result, #Correct Predicted/#Total represents accuracy when calculated. The numbers for true positives, false negatives, false positives, and true negatives are denoted by the letters TP, FN, FP, and TN, respectively.

(TP+TN)/(TP+TN+FP+FN) equals accuracy.

(21+28) / (21+28+9+3) = 0.80 = 80% accuracy

Predicting the Test set results

The first value is our estimated value, while the second is our actual value. If the data line up, our prediction was accurate. Given below the data visualization is mapped in a diagonal point. This diagonal point is used to identify the network path of the criminals Dark Web. Datasets named as VIDHARBA also mentioned below to identify the path as well as produce the accuracy in Machine learning Algorithms.

Data Visualization

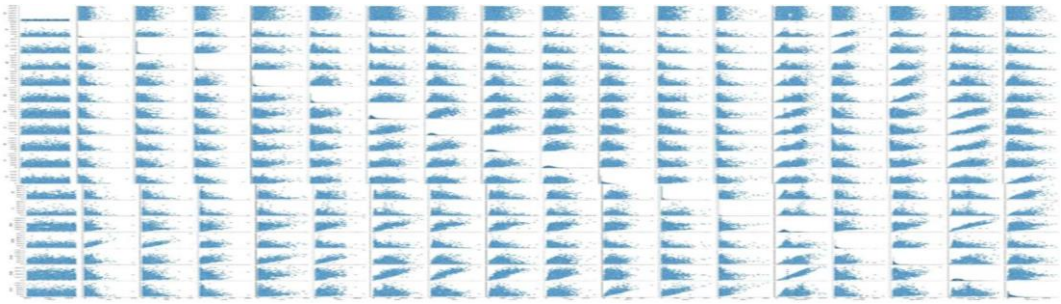

Figure 8 Data Visualization in Structure Patterns

Analysis of Dataset in VIDARBHA

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
2852	VIDARBHA	1901	36.8	39.9	30.9	26.1	7.3	129.7	295.3	368.8	123.4	35.2	0.0	0.0	1093.3	76.6	64.3	917.2	35.2
2853	VIDARBHA	1902	1.6	0.1	0.0	6.5	4.1	38.0	270.7	204.7	150.9	29.6	16.1	26.7	748.9	1.7	10.6	664.3	72.4
2854	VIDARBHA	1903	5.2	4.0	0.1	2.5	37.8	121.2	475.5	325.5	154.8	100.8	2.0	0.0	1229.4	9.3	40.3	1077.0	102.8
2855	VIDARBHA	1904	4.3	2.4	12.9	0.2	14.8	148.9	158.3	151.8	196.9	61.7	0.0	0.9	753.2	6.7	27.9	655.9	62.7
2856	VIDARBHA	1905	7.3	12.7	12.4	16.2	14.0	81.0	254.5	216.3	321.3	6.0	0.2	0.0	941.8	20.0	42.6	873.1	6.2

Table 1 Analysis of Dataset in VIDARBHA

Future Enhancement

There are several approaches to enhance this prediction system's scalability and accuracy. Now that we have a generalized framework, we can use it to analyze different types of data sets in the future. The performance of this prediction can be significantly improved by controlling multiple class labels during the prediction process, and this could be a productive area for further investigation. Since the crawling data in DM warehouses is usually highly dimensional, future research will likely have difficulty identifying and selecting critical attributes for more accurate data prediction.

Conclusion

Some people are still unaware of the Surface Web. Many times, people get confused between the web and the Internet. Actually, they are two different names that have certain things in common. The Internet consists of the extensive network infrastructure of multiple networks. It allows a million computers to be connected by creating a network where any computer can communicate with any other computer as long as it is connected to the Internet. Finally, we may draw a conclusion from this article using a recurrent neural network model that was created expressly for the setting of the surface web. This methodology is used to examine the illicit content that anonymous online crawlers find. Our RNN Model produces results with an accuracy of 80%. A decent accuracy rate is one that is above 70%.

References:

[1] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: applications, challenges and trends,” Neurocomputing, vol. 408, pp. 189–215, 2020.

[2] S. J. Devi and B. Singh, “Link prediction model based on the topological feature learning for complex networks,” Arabian Journal for Science and Engineering, vol. 45, no. 12, Article ID 10051, 2020.

[3] G. Wang, "Comparative study on different neural networks for network security situation prediction," Security and Privacy, vol. 4, no. 1, p. e138, 2021.
[4] S. Abdul, I. N. MamoonaQadir, M. A. Islam, and M. Aleem, "A comprehensive survey of link prediction techniques for social network," EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, vol. 7, no. 23, p. e3, 2020.
[5] J. Zhou, Y. Qiu, S. Zhu et al., "Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate," Engineering Applications of Artificial Intelligence, vol. 97, Article ID 104015, 2021.
[6] S. Kanmani and S. V. M. Satyanarayana, "Extremal states of qubit-qutrit system with maximally mixed marginals," Physics Letters A, vol. 385, Article ID 126978, 2021.
[7] A. S. Rajawat, P. Upadhyay, and A. Upadhyay, "Novel deep learning model for uncertainty prediction in mobile computing," Advances in Intelligent Systems and Computing, vol. 1, pp. 652–661, 2020.
[8] A. Singh Rajawat and S. Jain, "Fusion deep learning based on back propagation neural network for personalization," in Proceedings of the 2nd International Conference on Data, Engineering and Applications (IDEA), pp. 1–7, Bhopal, India, February, 2020.
[9] X.-W. Wang, Y. Chen, and Y.-Y. Liu, "Link prediction through deep generative model," iScience, vol. 23, no. 10, p. 101626, 2020.
[10] W. Feng, Z. Ma, R. Zhuang, and H. Che, "The framework of learnable kernel function and its application to dictionary learning of SPD data," Pattern Analysis and Applications, vol. 24, pp. 1–17, 2021.
[11] M. Yildirim, "Artificial intelligence-based solutions for cyber security problems," Artificial Intelligence Paradigms for Smart Cyber-Physical Systems, IGI Global, Hershey, PA, USA, pp. 68–86, 2021.
[12] M. J. Segovia-Vargas, "Money laundering and terrorism financing detection using neural networks and an abnormality indicator," Expert Systems with Applications, vol. 169, Article ID 114470, 2021.
[13] P. Haldar and A. Singh, "Machine intelligence versus terrorism," Artificial Intelligence and Global Society: Impact and Practices, pp. 147–159, 1st edition, 2021.
[14] M. T. Jafar, M. Al-Fawa'reh, Z. Al Hrahsheh, and S. Tayseer Jafar, "Analysis and investigation of malicious DNS queries using CIRA-CIC-DoHBrw-2020 dataset," Manchester Journal of Artificial Intelligence and Applied Sciences, vol. 26, 2021.