IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF MALICIOUS URLS IN BIG DATA USING RIPPER ALGORITHM

[1]SONIKA THAKUR
[1]STUDENT
[1]CENTRAL UNIVERSITY OF PUNJAB

Abstract—**'Big Data' is the term that describes a large amount of datasets. Datasets like web logs, call records, medical records, military surveillance, photography archives, etc. are often so large and complex, and as the data is stored in Big Data in the form of both structured and unstructured therefore, big data cannot be processed using database queries like SQL queries. In big data, malicious URLs have become a station for internet criminal activities such as drive- by-download, information warfare, spamming and phishing. Malicious URLs detection techniques can be classified into Non-Machine Learning (e.g. blacklisting) and Machine learning approach (e.g. data mining techniques). Data mining helps in the analysis of large and complex datasets in order to detect common patterns or learn new things. Big data is the collection of large and complex datasets and the processing of these datasets can be done either by using tool like Hadoop or data mining algorithms. Data mining techniques can generate classification models which is used to manage data, modelling of data that helps to make prediction about whether it is malicious or legitimate. In this paper analysis of RIPPER i.e. JRip data mining algorithm has been done using WEKA tool. A training dataset of 600 URLs has been made to train the JRip algorithm which is an implementation of RIPPER algorithm in WEKA. Training dataset will generate a model which is used to predict the testing dataset of 450 URLs. Accuracy are calculated after testing process. Result shows JRip has an accuracy of 82%.**

Keywords- Big Data, Data Mining, JRip, Weka, True positive rate, True negative rate, False positive rate, False negative rate, Accuracy.

I. INTRODUCTION

BIG DATA is the term that describes a large amount of datasets. Datasets like web logs, call records, medical records, military surveillance, photography archives, etc. are often so large and complex, and as the data is stored in Big Data in the form of both structured and unstructured therefore, big data cannot be processed using database queries like SQL queries [1]. The amount of data generated every day in the world is massive. The increasing volume of digital and social media and the internet of things is fueling it even further. The rate of data growth is surprising and this growth rate is really very fast, with variety (not necessarily structured) and contains a wealth of information that can be a key to gain the valuable knowledge in businesses. "Big data" is the term for a collection of data sets so large and complex that it becomes difficult to process it using traditional database management tools such as Relational Database Management System (RDBMS) [2]. RDBMS can't handle, huge, unstructured and complex data. The processing of large amount of dataset in RDBMS takes time as it is generally designed for fixed amount of data. So a different tools and techniques is needed to process the big datasets.

II. CHARACTERISTICS OF BIG DATA

Big Data is important because it enables organizations to gather, store, manage, and manipulate vast amounts of data at the right speed, at the right time and to generate the right result. Gartner defines big data in terms of 4Vs i.e. volume, variety, velocity and veracity or data quality [1]. Big data generators must create scalable data (Volume) of different types (Variety) under controllable generation rates (Velocity) while maintaining the important characteristics of the raw data (Veracity). Therefore, these four characteristics have been used to define Big Data and defined in [3].

III. APPLICATIONS OF BIG DATA

Big data is widely used in many areas. Some of these areas are as following [4]: -

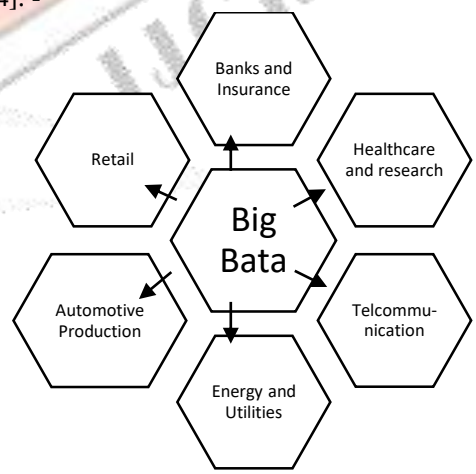

Figure 2: Applications of Big Data

A. *Banks and Insurance*

In Banks and Insurance big data is used to gather and process large amounts of data, for example customer details, policies, data entries, etc. The big data about customers and employees obtained from banks. Big data analysis can detect and prevent various types of frauds like insurance fraud, ATM card fraud etc. [4]

B. *Healthcare & Research*

In healthcare and research, big data has the possibility to make improvements in the quality of life. Information gained from big data in healthcare is used to increase the ability of patients to

monitor their own treatments and improve the ability of doctors to choose the best treatments for patents [4].

C. Telecommunication

Big data, which is gathered from the telephone users, can also help in mining for customer satisfaction, feedback, details, plotting charts, etc. from customer data, market research. Such data analysis helps in creating more services and offers to lure in customers [4].

D. Energy and Utilities

In this application big data is used to develop the smart meters. The smart meters provide a more accurate measure of energy usage by giving far more frequent reading than traditional meters. It gives several readings a day, not just once a month or once a quarter. Thus big data collection and processing is important in energy field also [4].

E. Automotive Production

Big data of automotive industry is applied to attract customers by adding features, designs, different insurances that can be beneficial for their automobile. Customer feedback, market shares, economic histograms can be used for analyzing different trends in business [4].

F. Retail

Retailers collect and maintain sales records for large number of customers [4]. A retailer would like to understand the characteristics of its customer. With big data, retailers can have instantly updated information about the size and location of inventories [4].

IV. MALICIOUS URLs

URLs have become a common channel to facilitate Internet criminal activities such as drive-by-download, spamming, phishing and information warfare. Many attackers use fake web sites for spreading malicious programs or stealing identities [1]. In this research data mining technique has been used to detect malicious URLs from the big data. Malicious URLs detection techniques [5] can be classified into Non-Machine Learning and Machine Learning approach.

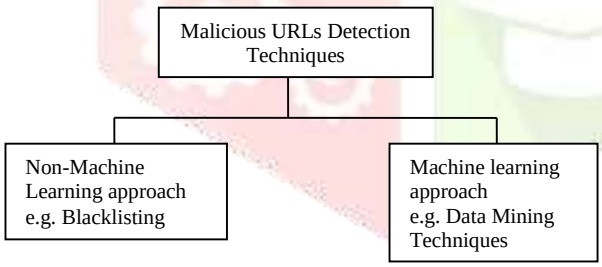

Figure 3: Malicious URLs Detection Technique

The non-machine learning approaches, suffer from poor generalization to new malicious URLs and unseen malicious patterns [5]. One of the examples of non-machine learning approach is blacklisting [6]. Many web browser uses blacklisting method (e.g. google, yahoo, etc.). Web browsers maintains a list of malicious URLs, this list is known as blacklist. Web browsers update their blacklist regularly. But some malicious URLs are online for only few hours. Because of this short life of malicious URLs, they may not get updated in blacklist and because of this web browsers may not be able to detect malicious URLs. This is the main disadvantage of blacklist approach [5].

Most used machine learning method is data mining techniques [5]. Data mining algorithms can be used to generate a classification model that can identify the malicious URLs. Malicious URLs generally have some common features like host name, favicon, etc. Data mining technique can utilize these feature in order to check URLs whether they are malicious or not. Data mining techniques generate a classification model from a training dataset (in this set URLs category are already defined i.e. malicious or legitimate) and then this model is applied on the testing dataset (in this set URLs category is not defined i.e. malicious or legitimate) to predict whether a URL is legitimate or malicious [5].

V. RIPPER ALGORITHM FOR MALICIOUS URLs DETECTION

RIPPER algorithm was designed by cohen in 1995 [7]. The RIPPER algorithm is one of the rule-based classification algorithm that generates rule-based classifier model which is a set of IF-THEN rules and these rules are extracted directly from the training dataset that's why it is called direct method. It is especially more efficient on large noisy datasets. The algorithm progresses through four phases [7]:

i. Growth: In the growth phase, one rule is generated by greedily adding attributes to the rule until the rule meets stopping criteria.

ii. Pruning: In the pruning phase, each rule is pruned and made shorter by removing redundancy and reducing length of earlier rule which allows the rule to become better.

iii. Optimization: The first growth and prune phase generates rules from empty rule set. Optimization step utilize the rules generated in first growth and prune stage and tries to generate new rules from ruleset. Rules is further optimized by:-
   - Adding attributes to the original rule using greedy method (i.e. depth first search).
   - New ruleset is generated after a growing and pruning phase.

iv. Selection: In the selection phase, the best rules are kept and the other rules are deleted from the model.

VI. METHODOLOGY

Data mining algorithm have been proved to be beneficial for detecting malicious URLs. After analyzing various URLs, their common features like host name, path length, etc. are extracted [8].

Waikato Environment for Knowledge Analysis (WEKA) tool has been used for the analysis of RIPPER algorithm [9]. JRip is an optimized implementation of RIPPER in WEKA tool which generates rulesets after the evaluation over the training dataset. The different steps involved in the analysis of JRip algorithm is as follows [9]:

i. Collection of both malicious URLs and legitimate URLs from Wiktionary_en_2012-07-21.hdt which is a big data database [10].

ii. Extract the features of URLs to detect the malicious and legitimate URLs.

iii. Creation of training dataset and testing dataset on the basis of features extracted.

iv. Training of RIPPER using training dataset and generation of rule-based classifier model.

v. Rule-based Classifier model is used to predict the missing values of testing dataset.

vi. URLs from testing dataset are predicted by using rule-based classifier model on the basis of different parameters such as True Positive Rate (TP rate), False Positive Rate (FP rate), True Negative (TN Rate), False Negative (FN Rate) and Accuracy.

A. Collection and extraction

Collection of both legitimate as well as malicious URLs is done for the extraction of their features. The Malicious as well as legitimate URLs are collected from "Wiktionary_en_2012-07-21.hdt". The extracted features of URLs are the basis for determining that whether a URLs is malicious or not. Twenty-five URL's features have been extracted which is used in training and testing dataset for categorizing the URLs [9].

B. Creating datasets

After extracting the features of various URLs, training and testing dataset is created. Both datasets are in .arff file format which is supported by WEKA [9]. Training and testing datasets are different from one another. In training dataset the category (i.e. malicious or legitimate) of URLs are already known. This category is represented as 'Result' attribute. Whereas in testing dataset the 'Result' attribute contains missing values.

A training dataset contains three parts [9]:

i. Relation: It is just a name for the training dataset.
ii. List of attribute: This contains the features of URLs and their values i.e. attribute value which can be either nominal or numeric
iii. Data: This contains the URLs' data according to the attributes.

Training dataset of 600 URLs has been created out of which 400 URLs are malicious and 200 URLs are legitimate.

Figure 4: Training dataset

C. Simulation Environment of Ripper (JRip) Algorithm

Table I shows the simulation environment to predict whether a URL in the testing dataset is malicious or legitimate using rule-based classifier model generated by JRip algorithm with training dataset:

Table I: Simulation Environment

S.no	Algorithm	Training dataset	Testing dataset	Parameters
1.	JRIP (RIPPER) Algorithm	Malicious URLs-400 Legitimate URLs-200	Malicious URLs-300 Legitimate URLs-150	TP Rate TN Rate FP Rate FN Rate Accuracy

i. Training Dataset: Training dataset is of legitimate as well as malicious URLs. Training dataset will contain the 200 legitimate URLs and 400 malicious URLs.
ii. Ripper algorithm will generate Rule-based classifier model.
iii. Testing Dataset: Testing dataset of 450 URLs is created in which 150 are legitimate URLs and 300are malicious URLs.

D. Preprocessing of training dataset and training of JRip algorithm

When a dataset is loaded in WEKA, it is first preprocessed. After preprocessing a list of all the attributes with their values is shown. For every attribute, calculation is done and number of instances having nominal value (like strings, characters) and numeric value (use of numbers i.e. 0 and 1) of attribute is determined.

JRip algorithm is selected in WEKA and it is trained with the training dataset. After this, the rule-based classifier model is saved for prediction of malicious as well as legitimate URLs. Ripper algorithm uses FOIL's information gain [7] in order to choose the attributes that can generate best rule. Every rules in RIPPER algorithm is created in the form of disjunctive normal form i.e. $R = r_1 \vee r_2 \vee \ldots .. r_k$, where R is rule set and r_i's are rules [7]. Attribute 'Favicon' is chosen first as it has highest FOIL's information gain. In the first rule attribute 'SSL' is added with and ('^') conjunction as it improves the rule's quality. Attribute 'SSL' is chosen as when it is added with attribute 'favicon' it gives larger FOIL's information gain. The other rules can be find by the same procedure [7].

After training of the JRip algorithm, the accuracy obtained for the algorithm is 94.833%. The 7 rulesets are generated by JRip.

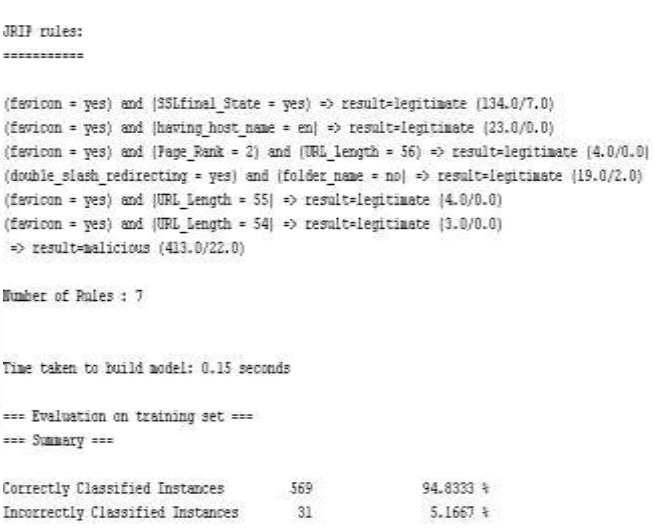

Figure 5: Rule-Based Classifier Model

a) A URL is legitimate if:

Rule 1: it has the value "yes" for both favicon and SSL final state.

Rule 2: it has the value "yes" for favicon and "en" for having_host_name.

Rule 3: it has the value "yes" for favicon and "2" for Page_Rank and "56 characters" value for URL_Length.

Rule 4: it has the value "yes" for double_slash_redirecting and "no" for folder_name. **Rule 5:** it has the value "yes" for favicon and "55 characters" value for URL_Length.

Rule 6: it has the value "yes" for favicon and "54 characters" value for URL_

b) A URL is Malicious if:

Rule 7: All the previous rules (from rule 1 to rule 6) are not satisfied by the URL then it will interpret that URL is malicious.

E. Testing of JRIP algorithm

The testing dataset is different from training dataset. In the testing dataset, the category (i.e. malicious and legitimate) of the URLs is unknown. It contains missing values. This missing attribute value is filled by the rule-based classifier which was saved as earlier. The testing dataset contains 450 URLs in which 300 are malicious URLs and 150 are legitimate URLs.

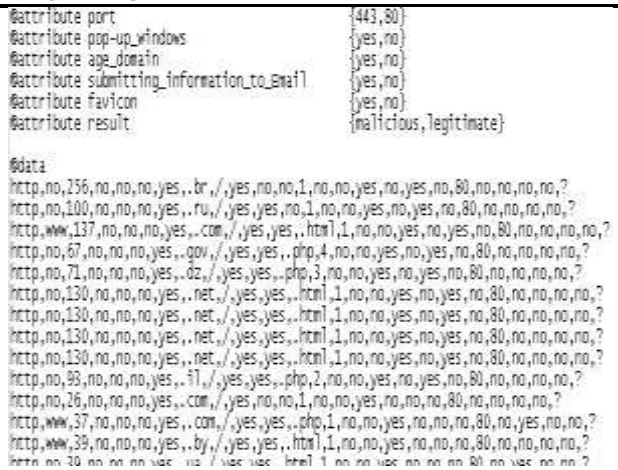

Figure 6: Testing dataset

Figure 6 shows the testing dataset in which the 'result' attribute is set missing (?) for every URL. The rule-based classifier model of JRip algorithm thus predicts the value of 'result' attribute [9].

Figure 7: Result of testing dataset

In WEKA from the "Test Options" tab, "Supplied test set" option is selected and testing dataset file is loaded. Finally, right click on the loaded model and run "Re-evaluate model on current test set". The results are shown in the "Classifier output" panel, under "Predictions on test data" [9] as shown in figure 7.

The symbol "+" occurs only for those items where error is encountered, that is the actual value for 'Result' attribute is different from its predicted value. As in testing dataset the values given for 'Result' attribute is "?" and after prediction 'Result' attribute has value either "malicious" or "legitimate", the symbol "+" occurs under the error column [9].

F. Result analysis

After the prediction of testing dataset using rule-based classifier model a confusion matrix is generated. The confusion matrix is useful for analyzing how well the rule-based classifiers can predict the unknown URLs. Each column of the confusion matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The true positive, true negative, false positive and false negative counts are used as a metrics and the performance of the algorithm's model is measured in terms of the accuracy.

The confusion matrix obtained is shown in table II.

Table II: Confusion matrix

	Predict malicious	Predict legitimate
Actual malicious	242 (True Positive)	58 (False Negative)
Actual legitimate	23 (False Positive)	127 (True Negative)

The different parameters are calculated from the confusion matrix. The above Table II has two rows and two columns that shows the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN).

The parameters that is to be calculated are explained below [11]:

- **True Positive (TP):** If the outcome from a prediction is malicious and the actual value is also malicious, then it is called a true positive

  True positive rate (TPR) =TP/ (TP+FN)

  TP is the True Positive, FN is the False Negative.
- **True negative (TN):** A true negative (TN) has occurred when both the prediction outcome and the actual value are legitimate.

  True negative rate is calculated as

  True Negative Rate =TN/ (TN+FN)

  TN is the True Negative, FN is the False Negative.
- **False Positive (FP):** False Positive (FP) is when the prediction outcome is malicious while the actual value is legitimate.

  False Positive Rate= FP/ (FP+TN)
- **False Negative (FN):** False negative (FN) is when the prediction outcome is legitimate while the actual value is malicious.

  False Negative Rate= FN/ (TP+FN)
- **Accuracy:** It is also referred as "correct classification rate" and is measured by taking the ratio of correctly prediction from the total URLs

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

Table III: Values of different parameters

Parameter Table for Ripper Algorithm		
Sl. No.	Parameters	Values
1	True Positive Rate	0.807
2	True Negative Rate	0.8466
3	False Positive Rate	0.153
4	False Negative Rate	0.1933
5	Accuracy	82%

Table III contains the values of parameters after calculation based on the formulas explained above. The value of different parameters are as follows: false positives (FP) is 0.153, false negatives (FN) is 0.1933, true positives (TP) is 0.807, and true negatives (TN) is 0.846, accuracy is 82%.

VII. CONCLUSION

Big data is a collection of large dataset. Big data is explained using four V's: volume, variety, velocity and veracity. In big data, malicious URLs have become a station for internet criminal activities such as drive- by-download, information warfare, spamming and phishing.

In this research work "Wiktionary_en_2012-07-21.hdt" big data database has been taken for creating training and testing datasets. The URLs' features are analyzed first in order to create training dataset. Twenty-five features of URLs are extracted for generating the training and testing dataset. A training dataset of 600 URLs and testing dataset of 450 URLs are created by using "Wiktionary_en_2012-07-21.hdt". The created training dataset is used to train the JRip algorithm in WEKA which generates a rule-based classifier model. The 450 URLs of testing dataset are predicted with rule-based classifier model, out of which 242

URLs are detected as malicious and 127 URLs are detected as legitimate. After this accuracy of the generated rule-based classifier model is calculated. The result shows that the rule-based classifier model of RIPPER algorithm can identify URLs with an accuracy of 82%.

VIII. REFERENCES

[1] C. Snijders, U. Matzat and U. D. Reips, ""Big Data" : Big Gaps of Knowledge in the Field of Internet Science," *International Journal of Internet Science,* vol. 7, no. 1, pp. 1-5, 2012.

[2] B. Due, M. Kristiansen and R. C. Palacios, "Introduction BIG Data Topics: A Multicourse Experience Report from Norway," in *Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality*, Porto, pp.565-569, 2015.

[3] . V. M. Schonberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think, Carolina: ISACA, 2014.

[4] M. Wessler, "Applications of Big Data," in *Big Data Analytics For Dummies*, Hoboken, Wiley, 2013, pp. 22-32.

[5] P. Zhao and S. C. Hoi, "Cost-Sensitive Online Active Learning with Application to Malicious URL Detection," in *ACM*, Chicago, 2013.

[6] J. Makey, "Blacklists Compared," 2010. [Online]. Available: https://www.sdsc.edu/~jeff/spam/Blacklists_Compared.html. [Accessed 17 august 2016].

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques, USA: Morgan Kaufmann, 2000.

[8] P. N. Tan, V. Kumar and M. Steinbach, Introduction to data mininig, Boston: pearson, 2006.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *Article*, vol. 11, no. 1, pp. 10-18, 2009.

[10] "Wikitionary_en_rdt.hdt," 21 july 2012. [Online]. Available: http://www.rdfhdt.org/datasets/. [Accessed 22 july 2016].

[11] I. H. witten and E. Frank, data mining practical machine learning tools and techniques, USA: Elsevier, 2005.