XXI INTERNATIONAL SCIENTIFIC-EXPERT CONFERENCE ON RAILWAYS
RAILCON '24
October 10-11, 2024; Niš, Serbia

Original scientific paper DOI: 10.5937/Railcon24087J

APPLICATION OF MACHINE LEARNING IN THE ANALYSIS OF THE WEB PROXY SERVER IN RAILWAY TRAFFIC ENVIRONMENT

Dragan JEVTIĆ [1]
Kristijan KUK [2] [0000-0001-8910-791X]
Vladica STOJANOVIĆ [3] [0000-0002-3819-4387]

Abstract – *The introduction of machine learning and anomaly detection classifiers in the process of supervision in the IT industry, opens up a new dimension in traffic analytics and decision-making for interventions in short periods. Case, described in his paper, needed a proxy server control for certain clients in the intranet network, clients in the traffic and transportation domain. The importance of those positions lies in the fact that the operation and control of traffic processes on the railway are monitored through these workstations. Since the National CERT of the Republic of Serbia issued a notice about possible threats via the Internet to users who have the Any Desk application on their workstations, the task was to analyze whether there is such a case with predefined workstations. The paper includes the analysis of event records from the proxy server using the Weka software package with the J48 classifier. The analysis included reports from the Weka program using a decision tree and the appropriate classifier.*

Keywords – WEKA, decision tree, proxy server, railway staff.

1. INTRODUCTION

Data analysis using classifiers and decision trees still has applications in machine learning, although new techniques are coming along over time. The work deals with the use of the J48 classifier in the analysis of data from the Proxy web server, in order to detect flaws in the blocking of the Any Desk software package on computers used by traffic officers, related to the text of the National CERT regarding the announcement that old versions of the software are not stable, i.e. they are dangerous to use.

2. SOFTWARE TOOLS

The software package Weka, which stands for "Waikato Environment for Knowledge Analysis", was used for data analysis and machine learning algorithms [1]. The software is defined as open-source and is under the GNU (General Public License) license, which enables the use, modification and sharing of the software with other users. Weka, as a software tool, has survived various additions and modifications, but still, in the world of machine learning, it occupies a position due to the easy and logical use of the software. Weka is a software tool whose function is data mining. To successfully search, that is, perform data mining, the software tool must have certain modules for preparation, classification, regression, grouping and visualisation of data. The input data set on which a type of machine learning is applied, as a result of which a decision tree is built, is called a training data set (eng. training set). Comparing different types of machine learning over the same training data set is used to determine learning performance. Algorithm J48 belongs to the algorithms that build a classification model from the data for training the system using decision trees. The algorithm strategy itself is top-down, meaning it defines the first attribute and then creates branches for each attribute value. The advantages of the decision tree are ease of understanding, simplicity, and working with different types of data [2-5]. The algorithm starts with the original node as the base node. At each iteration, the algorithm iterates through each unused attribute of the set and calculates the informational gain of that attribute. After that, it chooses the attribute with the lowest entropy and the highest value of information gain. Here entropia means a measure of uncertainty in a data set.

[1] Infrastructure of Serbian Railways, Nemanjina 6, Belgrade, Serbia, University of Criminal Investigation and Police Studies, Cara Dusana 196, Belgrade, Serbia, jevtic.dragan@gmail.com, dragan.jevtic@srbrail.rs
[2] University of Criminal Investigation and Police Studies, Cara Dusana 196, Belgrade, Serbia, kristijan.kuk@kpu.edu.rs
[3] University of Criminal Investigation and Police Studies, Cara Dusana 196, Belgrade, Serbia, vladica.stojanovic@kpu.edu.rs

The basic formula for entropy calculation (1) is:

$$H(S) = \sum_{x \in X} -p(x) log_2 p(x) \quad (1)$$

where S is the current data set over which entropy is calculated, X is the set of classes S, p(x) is the ratio of the number of elements in class x to the number of elements in set S.

Entropy, in information theory, measures how much information is expected to be it is obtained by measuring a random variable and as such can be used to quantify the quantity up to which the distribution of the value of the quantity is unknown. For example, a constant variable has zero entropy, while a uniformly distributed random variable has maximum entropy. The information gain measures the difference in entropy between the state before and after set S, which is divided into attribute A. The information gain is calculated according to the following formula (2):

$$IG(S,A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A) \quad (2)$$

where H(S) is the entropy of the set, T is the subset created by dividing the set into attributes, p(t) is the ratio of the number of elements in t to the number of elements in the set S, H(t) is the entropy of the subset t.

Evaluating the value of the classifier allows for defining the best classifier. Some of the methods used for this purpose are:

- Test sample method
- Method of cross-assessment

The test sample method divides the data set into two parts. The first part is the dataset that will be used for training, while the second will be used for testing. 33% of the total data is usually taken to assess the accuracy of the prediction. The accuracy score is a random number that directly depends on dividing the data set into training and test data sets. The procedure is repeated k times, and the accuracy can be obtained as the mean value. The prediction accuracy can be increased by increasing the number of samples in the training set. Figure 1 shows a model of data processing using model training and subsequent sampling with another set of data [4].

The quality of the qualification model cannot be defined by a single value. For this purpose, a graphical representation via the ROC (Receiver Operating Characteristic) curve is used. Sensitivity and specificity are two measures that characterise a classifier's specific accuracy.

The vertical axis of the ROC curve indicates the number of true positive samples, while the horizontal axis indicates the number of false positive samples, as shown in Figure 2 [6].

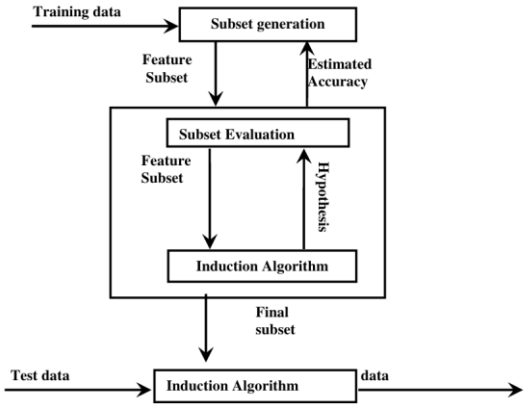

Fig. 1. A model of working with data [4]

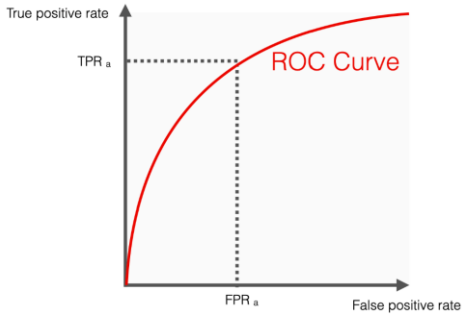

Fig. 2. ROC curve model [6]

3. DATA SET PREVIEW

The data used in the work were obtained from the central WEB Proxy server. The proxy server can centralise Internet access for workstations in the Intranet network. Each exit is recorded with a corresponding log and follows the server policies defined on the device and following the business policy [7,8]. In the example given in the paper, the operation of the new Proxy server was tested, but with the option that server policies were not activated. The proxy server is uploaded to the traffic dispatchers' workstations through the Microsoft Active Directory server policies. The corresponding figure for the system under analysis is given in Figure 3.

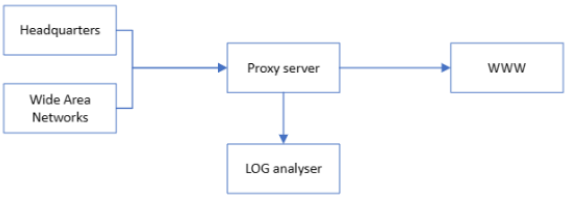

Fig. 3. Network model

For work and analysis of traffic to sites that are not allowed in the business environment, the following fields were selected as attributes for further analysis:

Duration is the time defined on the Proxy server and indicates the time it takes for the server to process the transaction session. This attribute can help us see the proxy server's usage for certain services. The following attribute values were defined for the analysis,

as shown in Table 1. For the sake of a better display and easier data processing, the classification was made into four groups in relation to the duration of the session.

Client IP is the client's address that initiates the session to the Internet. In our case, we look at two ranges of addresses, shown in Table 2. The basic ranges are workstations from the administration building and the range of workstations from the rest of the intranet network. The division was made based on the number of computers and sessions to facilitate later detection of problems in the system.

Tab. 1. View classified attribute Duration

Attribute lebel	Quantity	Value range
A	218005	0-1 msec
B	121598	1-500 msec
C	167779	500-4.000.000 msec
D	69317	100.000-4.000.000 msec

Tab. 2. Range view for client workstations (IP)

Attribute lebel	Quantity	Value range
HQ	281991	IP addresses of workstations located in the administrative building
WAN	225391	IP addresses of workstations located in the field, by railway stations and facilities

Tab. 3. Display of attribute values for the destination

Attribute lebel	Quantity	Value range
AnyDesk	11998	Remote computer access
Facebook	6363	Social network
Google	38078	Google services
Microsoft	100842	Microsoft services
Ostalo	350101	Other services

Tab. 4. Display of attributes for session type

Attribute lebel	Quantity	Value range
App	3064	Defined applications used for Office business
Media	82210	Audio and Video Production
P2P	179183	Direct communication and data exchange between server and client
Tekst	242925	Textual content

Tab. 5. Display of session result attributes

Attribute lebel	Quantity	Value range
Cache	91311	Content is cached on the server
Denied	239448	Content has been rejected
Tunnel	176623	Content is tunnelled

URL (Uniform Resource Locator)—is an attribute that indicates the destination address of the client request. For the purposes of the analysis, the URL attribute values were defined as shown in Table 3. The emphasis is on the application that was detected as problematic.

Type—is an attribute that more closely determines the type of content exchanged between the client and the server. Considering the problem, all data are grouped as indicated in Table 4.

The resulting code is an attribute that gives the exact output model, indicating whether the session is rejected, whether the traffic is tunnelled directly to the server, or whether the content is taken from the local cache. The attribute values are shown in Table 5.

4. MODEL ANALYSIS

For the analysis of the data set, as stated, we will use the software package Weka, as well as the corresponding algorithm J48, through the analysis of the decision tree. Given that we have a data set of 507382 records, we will divide it into two sets; that is, we will define within the Weka software the division into 60% of the data set for training and 40% of the data set for test, as shown in Figure 4.

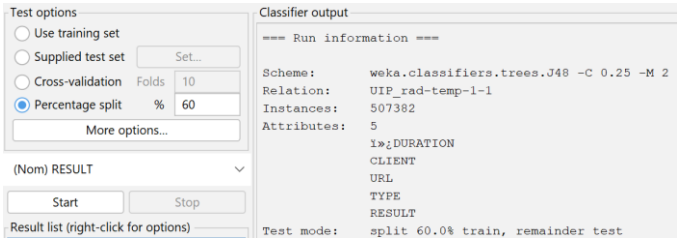

Fig. 4. Representation of the data set partitioning.

By running the J48 algorithm through the defined 60/40 data division, we get the following results in sum:

- 201177 correctly classified instances, about 99.1249% of the total test sets.
- Incorrectly classified instances 1776, about 0.8751% of the total test sets.

The confusion matrix is shown in Table 6. Other data, for deeper model analysis, for the J48 classifier are given in Table 7.

The accuracy classification of the model is confirmed by the ROC (Receiver Operating Characteristic Curve) Area values, which in our case is quite high. This means that the model can make predictions in a very high percentage. The graphical indicator of accuracy is the area under the ROC curve. The display of the ROC curve for the attribute Results, value Denied, is given in Figure 5.

After inspecting the decision tree, it can be determined that, in the interesting example, an application that is not allowed is detected in the tree and

blocked, as shown in Figure 6.

Tab. 6. Elements of confusion matrix

a	b	c	
35321	785	402	a = Cashe
219	95115	368	b = Denied
2	0	70741	c = Tunnel

Tab. 7. Overview of model values with classifier J48

TP Rate	FP Rate	Precision	Recall	Class
0.967	0.001	0.994	0.967	Cashe
0.994	0.007	0.992	0.994	Denied
1.000	0.006	0.989	1.000	Tunnel
0.991	0.006	0.991	0.991	Avg.
F-Measure	**MCC**	**ROC Area**	**PRC Area**	**Class**
0.980	0.976	0.995	0.989	Cashe
0.993	0.986	0.998	0.997	Denied
0.995	0.992	0.998	0.992	Tunnel
0.991	0.986	0.997	0.994	Avg.

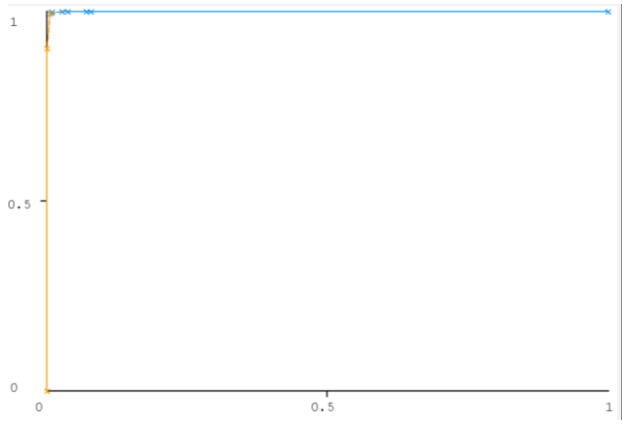

Fig. 5. ROC curve for the value of the Denied attribute Results

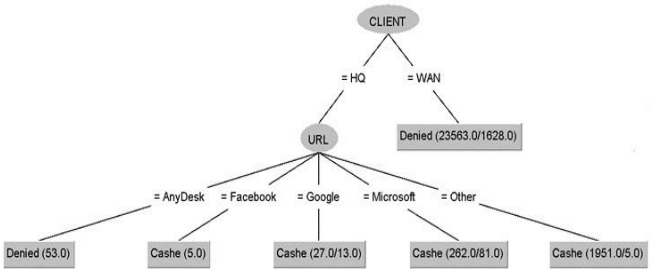

Fig. 6. Rejection of AnyDesk instance URL attribute

5. CONCLUSION

This paper disscused the principle of the Proxy server analysis and the possibility of events prediction using the Weka software package for machine learning and the J48 classifier were reviewed. The data used are real data, which do not often contain some undetermined values. In this case, we have mostly good data because the values for poorly classified values are around 0.875%, which is not a high value. The accuracy of well-classified data is 99.1249%, which is an extremely high percentage. By showing the confusion matrix and calculating the corresponding variables, TP, FP, TN and FN, we get that the model's accuracy is 0.991 as an average value, looking at the values of the Results attribute. Also, the value for ROC is 0.997, which represents a high value. By displaying the decision tree, it can be seen that the functionality of the Proxy server is completely mapped onto the tree. The main function was to predict and prove that certain applications, which are prohibited for certain employees of the traffic service, cannot be used. The tree clearly shows that those requests to use the prohibited application were rejected.

The next step in event prediction research could be choosing other classifiers, comparing different classifiers and selecting those whose accuracy would be higher than the one currently processed. Also, instead of one data set, it is preferable to use two data sets, one for training and the other for testing. The last type of analysis would be the adjustment of the model when the data sets were taken from different time intervals and for different days during the working week.

REFERENCES

[1] Eibe Frank, Mark A. Hall, and Ian H. Witten The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[2] Aleksa Maksimović, „Application of the decision tree in the analysis of samples and types of traffic accidents in the territory of the Republic of Serbia ", Belgrade, 2022

[3] Branislava Cvijetić, Zaharije Radivojević, " Application of Weka software and convolutional neural networks in the process of classifying digital photos ", , INFOTEH-JAHORINA, 2019

[4] Amira Abdalla, Sarina Sulaiman, Waleed Ali „Intelligent Web Objects Prediction Approach in Web Proxy Cache Using Supervised Machine Learning and Feature Selection", Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015

[5] Marko Mićalović, Gabrijela Dimić, Dragana Prokin, Kristijan Kuk, " Application of Decision Trees and Naïve Bayes classifiers to a dataset extracted from a Moodle course ", INFOTEH-JAHORINA, 2012

[6] Evidently AI, web: https://www.evidentlyai.com/

[7] Abdul Samed Isman, Waleed Ali, Sisi Mariyam, "Patterns Analysis and Classification for Web Proxy Cache", International Conference on Hybrid Intelligent Systems, 2011

[8] Jon Laurence, Rhonalda Wenceslao, "Network Performance of Proxy-Enabled Server using Three Configurations", Indian Journal of Science and Technologies, 2021