METHODS OF THE DATA MINING AND MACHINE LEARNING IN COMPUTER SECURITY

Norbert ÁDÁM, Branislav MADOŠ, Marek ČAJKOVSKÝ, Ján HURTUK, Tomáš TOMČÁK
Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, tel. +421 55 602 3023,
e-mail: {norbert.adam, branislav.mados, marek.cajkovsky, jan.hurtuk}@tuke.sk, tomas.tomcak@student.tuke.sk

ABSTRACT

Due to the expansion of high-speed Internet access, the need for secure and reliable networks has become more critical. Since attacks are becoming more sophisticated and networks are becoming larger there is a need for an efficient intrusion detection systems (IDSs) that can distinguish between legitimate and illegitimate traffic and be able to signal attacks in real time, before serious damages are produced. Although there are some existing mechanisms for intrusion detection, there is need to improve the performance. Data mining techniques and machine learning are a new approach for intrusion detection. In this work naive Bayes classifier and decision trees with C4.5 and CART algorithms for detecting abnormal traffic patterns in the KDD Cup 1999 data are used. The IDS system is supposed to distinguish normal traffic from intrusions and to classify the intrusions into five classes: Normal, DoS, probe, R2L and U2R. Research shows that decision trees gives better overall performance than the naive Bayes classifier.

Keywords: *Data mining, machine learning, classification, decision tree, naive Bayes classifier, computer security, intrusion detection system (IDS)*

1. INTRODUCTION

The Internet has grown tremendously in recent decades. The interconnection of computers and network devices has made the cyberspace so complex that even the best experts on the planet do not fully understand its deepest inner workings. Personal computers get faster every year and it is not rare for ordinary user to connect to the internet through 10 Mb/s lines or faster. To complicate the matter, there is an increase of the number and level of confidentiality of the information found on the internet. On-line banking and online payment make life easier for average internet user, but make it harder for security experts and network administrators. Practically anyone has access to the huge database of information that is the internet. In particular, many websites display information about software security and hacking. Others provide hacking toolkits that even the most inexperienced users can launch without difficulty. Add to this the fact that the Internet was not built in the interest of security because nobody could have predicted its dazzling expansion and we have the ingredients of a monumental problem.

Due to the availability of large amounts of data in cyber infrastructure and the number of cyber criminals attempting to gain access to the data, data mining, machine learning, statistics, and other interdisciplinary capabilities are needed to address the challenges of cybersecurity. Data mining is the extraction, or mining, of knowledge from a large amount of data. The strong patterns or rules detected by data-mining techniques can be used for the nontrivial prediction of new data. In nontrivial prediction, information that is implicitly presented in the data, but was previously unknown is discovered. Data mining uses statistical models, mathematical algorithms, and machine learning methods to discover previously unknown, valid patterns and relationships in large data sets, which are useful for finding attacks [1] [5][6][7].

2. GOALS

The goal of this paper is to build a predictive model (a classifier) capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections with Data mining and machine learning techniques. Proposal of the process can be seen on the Figure 1.

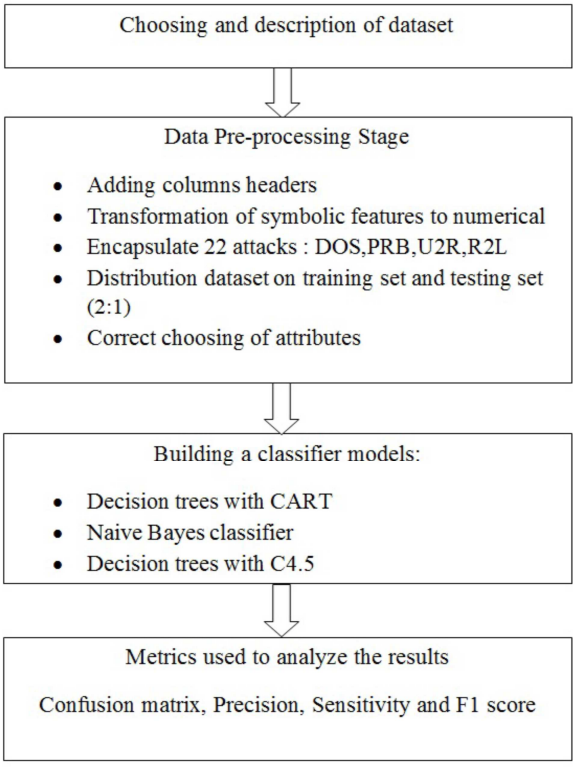

Fig. 1 Proposal of the process.

2.1. Dataset

In this work was used 10 percent KDD 1999 cup data set. The 10 percent KDD cup 1999 contains 494020 entries. Each instance in the KDD Cup 1999 datasets contains 41 features that describe a connection and a label (information about type of attack). Features 1-9 stands for the basic features of a packet, 10-22 for content features, 23-31 for traffic features and 32-41 for host based features [4]. There are 22 different attack types in dataset 10 percent KDD and these attack types fall into four main categories: probe, denial of service (DoS), remote to local (R2L) and user to root (U2R).

We divided 10 percent KDD in to training and testing sets(2:1). Models from training set used for testing set and sample Corrected. It is important to note that sample Corrected is not from the same probability distribution as the training data, and it includes specific 14 attack types not in the training data. This makes the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the "signature" of known attacks can be sufficient to catch novel variants [4] After basic statistical calculations two main problems were found:

1. The dataset is highly unbalanced. In particular, the classes U2R and R2L are the least well represented with only 52 and 1,126 examples respectively, whereas the DoS class contains 391458 examples. With such a small number of examples as in the U2R and R2L classes, it can be expected that it will be difficult for the classifier to predict the correct classes of unseen examples (Table 1).
2. The dataset has a lot of attributes.

Table 1 Statistics of various datasets.

Category	DOS	NORMAL	PROBE	R2L	U2R
10% sample	391458	97277	4107	1126	52
Testing set	129032	31984	1326	361	18
Training set	262426	65293	2781	765	34
Corrected	229875	60592	4166	16327	68

2.2. Data Pre-processing Stage

Adding columns headers: Since the KDD Cup 1999 dataset was retrieved unlabeled, one of the first important steps is to add columns headers to it. 41 columns headers are added that contain information such as duration, protocol type, service, src bytes, dst bytes, ag, land, wrong fragment , etc.

Transformation of symbolic features to numerical: The 10 percent KDD99 benchmark dataset has three symbolic features: protocol type, service and ag. These features are very important and should not be ignored.

Encapsulate 22 attacks (making target attribute): The next step is to encapsulate the attack names to their categories, 22 attack names to their four original categories DoS, PRB, R2L and U2L and normal to normal category. For the solution of first problem are used oversampling minority classes and undersampling majority classes.

Balanced data set:

- DOS 49,3 percent against 79,21
- Normal 24,65 percent against 19,71
- Probe 20,25 percent against 0,84
- R2L 5,55 percent against 0,23
- U2R 0,25percent against 0,01

Balanced data set did not have expected results for categories R2L and U2R. In this case weights were used. Another approach to make decision trees more suitable for learning from extremely imbalanced data follows the idea of cost sensitive learning. Since the decision trees classifier tends to be biased towards the majority class, we shall place a heavier penalty on misclassifying the minority class. Reversed weights in inverse proportion to their occurrence were used. Biggest weights have minority classes and smallest weights have majority classes. This solution gave us better results for Probe, R2L and U2R.

Correct choosing attributes: For the solution of the second problem several approaches were used (functions from R like csf, consistency and random forest). Function cfs is based on correlation and entropy. Function consistency is capable of finding attribute subset using consistency measure. The variable importance plot is a critical output of the random forest algorithm. For each variable in your matrix it tells you how important that variable is in classifying data. The plot shows each variable on the y axis, and their importance on the x axis. They are ordered top to bottom as most to least important. Therefore, the most important variables are at the top and an estimate of their importance is given by the position of the dot on the x axis.

We can see the most important attributes in Figure 2 and Figure 3. After several tests were used these attributes:

- For decision trees with algorithms CART and C4.5 : *wrongfragment* + *loggedin* + *srcbytes* + *count*+ *dstbytes* + *dsthostsrvdifhostrate*+ *isguestlogin*+ *hot* + *ag* + *dsthostsamesrcportrate* + *numcompromised* + *srvd-ifhostrate*+*srvcount* + *dsthostserrorrate*+ *di_srvrate* + *service* + *protocoltype* + *dsthostrerrorrate* + *duration*

- For Naive Bayes classifier*: dsthostsamesrvrate* +*dsthostsamesrcportrate dsthostsrvcount* + *dsthostcount* + *samesrvrate* + *srvcount* + *count* + *dstbytes* + *srcbytes* + *ag* + *service* + *duration* + *protocoltype* + *numfilecreations*+*loggedin* + *hot*

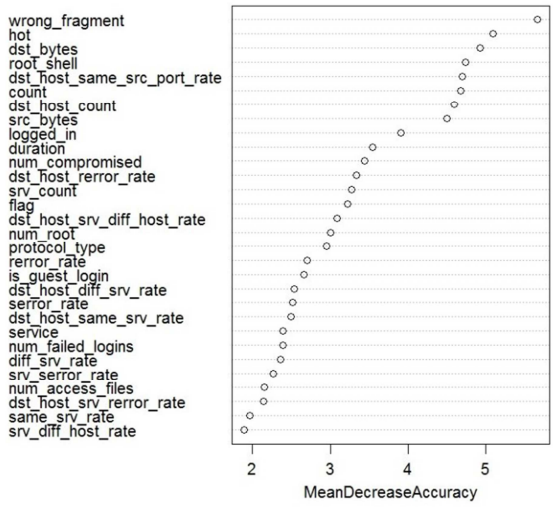

Fig. 2 Important attributes.

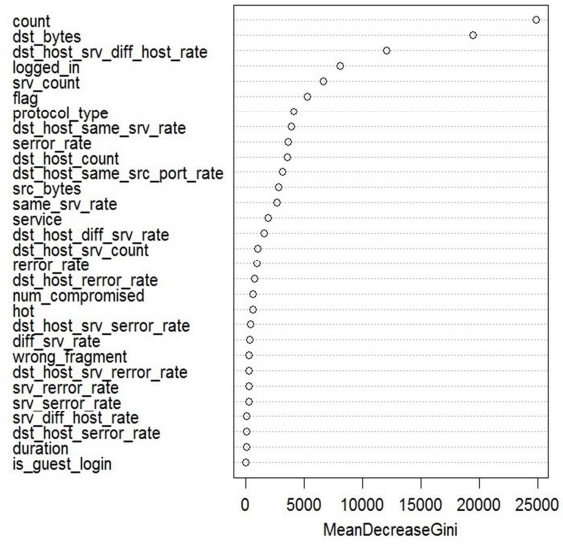

Fig. 3 Important attributes.

3. BUILDING A CLASSIFIER MODELS

We can divide classification in two steps [3]:

1. Building the classifier or model. This step is the learning step or the learning phase. In this step the classification algorithms build the classifier. The classifier is built from the training set made up of database tuples and their associated class labels. Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points. Building model with decision trees-algorithms for constructing decision trees usually work top-down, by choosing an attribute at each step that best splits the set of items. Different algorithms use different metrics for measuring "best". Algorithm CART (classification and regression tree) uses Gini Index and algorithm C4.5 uses information gain. Building model with Naive Bayes classifier computes the conditional aposterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule.

2. Using classifier for classification - In this step the classifier is used for classification. Here the test data are used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

4. EVALUATION CRITERIA (METRICS USED TO ANALYZE THE RESULTS)

Several metrics are used to evaluate and compare the performance of Intrusion Detection Systems (IDSs). The most basic metrics are Confusion matrix, Precision, Sensitivity and F1 score [2] (Figure 4).

- False Positive (FP): represents the number of instances that are classified by the intrusion detection systems as being anomalous when in fact they are legitimate.

- True Positive (TP): represents the number of instances that are classified by the intrusion detection systems as being anomalous and that really are anomalous.

- False Negative (FN): represents the number of instances that are classified by the intrusion detection systems as being legitimate when in fact, they are anomalous.

- True Negative (TN): represents the number of instances that are classified by the intrusion detection systems as being legitimate and that really are legitimate.

		Actual class	
		Negative class (Normal)	Positive class (Attack)
Predicted class	Negative class (Normal)	True Negative (TN)	False Negative (FN)
	Positive class (Attack)	False Positive (FP)	True Positive (TP)

Precision(P): $\frac{TP}{(TP+FP)} \times 100$

Sensitivity(R): $\frac{TP}{(TP+FN)} \times 100$

F1 score: $\frac{2 \times P \times R}{P+R}$

Fig. 4 Confusion matrix, Precision, Sensitivity, F1 score.

Table 1 Comparison of models before and after, sample corrected.

	Sample Corrected						
	DOS	NORMAL	PROBE	R2L	U2R	ACC	
CART	98,46	82,93	66,93	0	0	91,90%	
NBK	83,42	72,96	15,28	2,1	1,8	73,05%	
J48	98,53	83,98	81,94	5,63	29,55	92,42%	
	Sample Corrected KDD						
	DOS	NORMAL	PROBE	R2L	U2R	ACC	Editing
CART	98,54	84,99	82,93	7,26	28,26	92,64%	CP=0,00001 + attributes
NBK	92,47	74,12	32,6	0,05	3,1	88,52%	attributes
J48	98,7	86,69	66,44	11,45	32, 26	92,99%	U=TRUE+ attributes

Table 2 Comparison of CART without and with reversed weights, testing set and sample.

	Testing set 10 % KDD					
	DOS	NORMAL	PROBE	R2L	U2R	ACC
CART	99.59	96,22	69,14	24,61	23,63	97.92%
	Sample Corrected					
CART	98,1	83,33	71,69	19,47	26,28	91,67%

5. CONCLUSION

The goal of this work is to build a predictive model (a classifier) capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections with data mining and machine learning techniques. In this work were used naive Bayes classifier and decision trees with C4.5 and CART algorithms for detecting abnormal traffic patterns in the KDD Cup 1999 data. We used open source program project R. Models from training set were used for testing set and sample Corrected.

It is important to note that sample Corrected is not from the same probability distribution as the training data, and it includes specific 14 attack types not in the training data. First models were created without setup of parameters and with all attributes. Then it was possible to compare the performance of these models with final models.

After basic statistical calculations two main problems were found. First problem was the highly unbalanced dataset. This problem was solved by reversed weights in inverse proportion to their occurrence. Performance of Probe was better on 4,76 percent, R2L from 0 percent to 19,47 percent and U2R from 0 percent to 26,28 percent. Second problem was in higher number of attributes that were slowing down calculations. This problem was solved by several techniques for finding the best attributes. This step improved performance and time for calculations. With correct choosing of attributes, setup parameters and weights performance was significantly improved for all models. Comparison of the performance can be seen in Table 1 and Table 2.

Research shows that decision trees give better overall performance than the naive Bayes classifier.

ACKNOWLEDGMENTS

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0008-10 and project KEGA 008TUKE-4/2013: Microlearning environment for education of information security specialists.

REFERENCES

[1] DUA, S. – DU, X. : Data Mining and Machine Lerning in Cybersecurity. :CRC PRESS, 2011, 248 s. ISBN-13: 978-1439839423.

[2] RICHTER, J.: Dolování z dat v oblasti počítačové bezpečnosti : Koncept dizertačnej práce. Brno:VUT, 2011. 31 s.

[3] PARALIČ, J. : Objavovanie znalostí [online].Košice: TU, FEI,. [s.a.]. [cit. 2014-01-02]. Available on Internet: <http://people.tuke.sk/jan.paralic/prezentacie/OZ/Klasi_kacia.pdf> .

[4] The UCI KDD Archive Information and Computer Science University of California, Irvine Irvine, CA 92697-3425: 1999 – Available on Internet: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[5] WITTEN, I. H. – FRANK, E.: Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, June 2005, ISBN 0-12-088407-0.

[6] VOKOROKOS, L. – BALÁŽ, A. – TRELOVÁ, J. – ŠESTINA, J. – TURINSKÁ, A.: Protocol Intrusion Detection Architecture, In: CNSCE 2014 : International Conference on Computer, Network Security and Communication Engineering : Shenzhen, China, Februar 22-23, 2014, Lancaster : DEStech Publications, 2014, pp. 246-250, ISBN 978-1-60595-167-6.

[7] VOKOROKOS, L. – FANFARA, P. – RADUŠOVSKÝ, P. – POÓR, P: Sophisticated Honeypot mechanism – the autonomous hybrid solution for enhancing computer system security, In: SAMI 2013 : IEEE 11th International Symposium on Applied Machine Intelligence and Informatics : proceedings : January 31 – February 2, 2013, Herľany, Slovakia – Budapest: IEEE, 2013, pp. 41-46 - ISBN 978-1-4673-5926-9.

Received February 11, 2014, accepted June 24, 2014

BIOGRAPHIES

Norbert Ádám (Ing., PhD.) was born on 30.8.1980. In 2003 he graduated (MSc.) with distinction at the Department of Computers and Informatics at the Faculty of Electrical Engineering and Informatics of the Technical

University of Košice. He defended his PhD. in the field of Computers and computer systems in 2007; his thesis title was "Contribution to simulation of feed-forward neural networks on parallel computer architectures". Since 2006 he is working as a professor assistant at the Department of Computers and Informatics. Since 2008 he is the head of the Computer Architectures and Security Lab. at the Department of Computers and Informatics. His scientific research is focused on the parallel computers architectures.

Branislav Madoš (Ing., PhD.) was born on 20.5. 1976, in Trebišov, Slovakia. In 2006 he graduated (MSc.) with distinction at the Department of Computers and Informatics at the Faculty of Electrical Engineering and Informatics of the Technical University of Košice. He defended his PhD. in the field of Computers and computer systems in 2009; his thesis title was "Specialized architecture of data flow computer". Since 2010 he has been working as a professor assistant at the Department of Computers and Informatics. His scientific research is focused on the parallel computer architectures and architectures of computers with data driven computational model.

Marek Čajkovský (Ing.) was born on 17th December 1986 in Veľký Krtíš, Slovakia. In 2011 he graduated (MSc.) at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice and received the engineering degree. Since 2011 he is PhD. student at Faculty of Electrical Engineering and Informatics at Technical University of Košice. His research is focused on computer security, the title of his doctoral thesis is: Identifying Security Threats by System Services Calling. His professional interests include programming, computer networking, computer security and UNIX based operating systems.

Ján Hurtuk (Ing.) was born on 4th October 1988 in Kežmarok. In 2013 he graduated (MSc.) at the Department of Computers and Informatics at the Faculty of Electrical Engineering and Informatics of the Technical University of Košice. Since 2014 he is studying as a PhD. student at the Department of Computers and Informatics at the Faculty of Electrical Engineering and Informatics of the Technical University of Košice. His scientific research is mainly focused on the computer security.