ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ

әл-Фараби атындағы Қазақ ұлттық университетінің

ХАБАРЛАРЫ

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ
НАУК РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный
университет имени аль-Фараби

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
al-Farabi Kazakh National University

SERIES
PHYSICS AND INFORMATION TECHNOLOGY

2 (346)

APRIL – JUNE 2023

PUBLISHED SINCE JANUARY 1963

PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика және информатика сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.

Тақырыптық бағыты: *физика және ақпараттық коммуникациялық технологиялар сериясы.*

Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

Тиражы: *300 дана.*

Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Типографияның мекен-жайы: «Аруна» ЖК, Алматы қ., Мұратбаев көш., 75.

«Известия НАН РК. Серия физика и информатики».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика и информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75.

News of the National Academy of Sciences of the Republic of Kazakhstan.
Series of physics and informatics.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *series physics and information technology.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Address of printing house: ST «Aruna», 75, Muratbayev str, Almaty.

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 2. Namber 346 (2023). 21–31
https://doi.org/10.32014/2023.2518-1726.181

UDC: 004.49
IRSTI: 28.23.25

© G. Alpysbay[1*], A. Bedelbayev[1], O. Ussatova[1,2], A. Zhumabekova[1], Edzard Höfig[3], 2023
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Institute of Information and Computational Technologies, Almaty, Kazakhstan;
[3]Berlin University of Applied Sciences and Technology, Berlin, Germany.
E-mail: gulbanu.alpysbay@gmail.com

APPLICATION OF MACHINE LEARNING ALGORITHM IN THE ANALYSIS OF MALICIOUS SOFTWARE

Alpysbay G.E. — Lecturer. Al-Farabi Kazakh National University. Department of Information technology. 050040. Almaty, Kazakhstan
E-mail: gulbanu.alpysbay@gmail.com. ORCID: 0000-0002-3766-2396;
Bedelbayev A.A. — Candidate of science (physical and mathematical), professor. Al-Farabi Kazakh National University. Department of Information technology. 050040. Almaty, Kazakhstan
E-mail: agyn08@yandex.ru. ORCID: 0000-0001-9839-4156;
Ussatova O.A. — PhD, chief scientific secretary, s enior researcher. I nstitute o f I nformation and Computational Technologies. 050010. Almaty, Kazakhstan. PhD, acting associate professor. Al-Farabi Kazakh National University. Department of Information technology. 050040. Almaty, Kazakhstan
E-mail: uoa_olga@mail.ru. ORCID: 0000-0002-5276-6118;
Zhumabekova A.T. — Lecturer. Al-Farabi Kazakh National University. Department of Information technology. 050040. Almaty, Kazakhstan
E-mail: zhumabekova2702@gmail.com. ORCID: 0000-0003-4242-7988;
Edzard Höfig — Prof. Dr.-Ing., Berlin University of Applied Sciences and Technology. Department of Media and Computer Science, Berlin, Germany
E-mail: edzard.hoefig@bht-berlin.de.

Abstract. Rapid development and widespread use of information technologies and the Internet have many benefits, it is important not to ignore the negative situations that occur there. One of these negative situations is the proliferation and development of malicious programs, which lead to the failure of many devices and software equipment, and the disclosure of confidential information. Malware is the vehicle for many computer attacks and security breaches. Malware analysis uses techniques from several different fields, such as program analysis and network analysis, to study malicious patterns to gain a deeper understanding of several aspects, including their behavior and how they evolve over time. The day-by-day development of malicious software equipment, the increase in types and complexity of their structure has greatly complicated the work of identifying and classifying

them. For this reason, the importance of using artificial intelligence systems, including machine learning algorithms, is increasing in the field of information protection. Machine learning is a branch of artificial intelligence that involves algorithms and processes that "learn" and can generalize past data and insights to predict future outcomes. Basically, machine learning is a set of mathematical methods implemented in computer systems that provide the process of obtaining information, identifying patterns, and drawing conclusions from data. The article analyzes the type of malicious software in a certain format and classifies it with the help of a machine learning algorithm based on specific selected features. Here, the role of the selection of received features for training is very important. Machine learning works well when the input data is good.

Keywords: information security, malware analysis, machine learning, feature selection, kNN

© Г. Алпысбай[1*], А. Бедельбаев[1], О. Усатова[1,2], А. Жұмабекова[1], Эдзард Хофиг[3], 2023

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;

[2]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;

[3]Берлин қолданбалы ғылымдар және технологиялар университеті, Берлин, Германия.

E-mail: gulbanu.alpysbay@gmail.com

ЗИЯНДЫ БАҒДАРЛАМАЛЫҚ ЖАБДЫҚТАРДЫ ТАЛДАУДА МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМІН ҚОЛДАНУ

Алпысбай Г.Е. — Оқытушы. Әл-Фараби атындағы ҚазҰУ. Ақпараттық технологиялар факультеті. 050040. Алматы, Қазақстан
E-mail: gulbanu.alpysbay@gmail.com. ORCID: 0000-0002-3766-2396;
Беделбаев А.А. — Физика-математика ғылымдарының кандидаты, профессор. Әл-Фараби атындағы ҚазҰУ. Ақпараттық технологиялар факультеті. 050040. Алматы, Қазақстан
E-mail: agyn08@yandex.ru. ORCID: 0000-0001-9839-4156;
Усатова О.А. — PhD, бас ғылыми хатшы, аға ғылыми қызметкер. Ақпараттық және есептеу технологиялары институты. 050010. Алматы, Қазақстан. PhD, доцент м.а. Әл-Фараби атындағы ҚазҰУ. Ақпараттық технологиялар факультеті. 050040. Алматы, Қазақстан
E-mail: uoa_olga@mail.ru. ORCID: 0000-0002-5276-6118;
Жұмабекова А.Т. — Оқытушы. Әл-Фараби атындағы ҚазҰУ. Ақпараттық технологиялар факультеті. 050040. Алматы, Қазақстан
E-mail: zhumabekova2702@gmail.com. ORCID: 0000-0003-4242-7988;
Эдзард Хофиг — Берлин қолданбалы ғылымдар және технологиялар университеті профессоры. Медиа және информатика кафедрасы, Берлин, Германия
E-mail: edzard.hoefig@bht-berlin.de.

Аннотация. Ақпараттық технологиялар мен интернет желісінің қарқынды дамуы мен кеңінен қолданылуының тиімді тұстары көп болғанымен, ондағы орын алатын келеңсіз жағдайларды назардан тыс қалдырмау керек. Осындай келеңсіз жағдайлардың бірі — зиянды бағдарламалардың көбеюі және

дамуы, бұл көптеген құрылғылар мен бағдарламалық құрал-жабдықтардың істен шығуына, құпия ақпараттың ашылуына әкеліп соғуда. Зиянды бағдарлама көптеген компьютерлік шабуылдар мен қауіпсіздікті бұзудың құралы болып табылады. Зиянды бағдарламалық құралды талдау бірнеше аспектілерді, соның ішінде олардың мінез-құлқын және уақыт өте келе қалай дамып жатқанын тереңірек түсіну үшін зиянды үлгілерді зерттеу үшін бағдарламаларды талдау және желіні талдау сияқты бірнеше түрлі салалардағы әдістерді пайдаланады. Зиянды бағдарламалық жабдықтың күн санап дамуы, олардың түрлерінің ұлғаюы және құрылымының күрделілігі оларды анықтау және жіктеу жұмысын айтарлықтай қиындатады. Осы себепті ақпаратты қорғау саласында жасанды интеллект жүйелерін, соның ішінде машиналық оқыту алгоритмдерін пайдаланудың маңыздылығы артып келеді. Машиналық оқыту — бұл «үйренетін» алгоритмдер мен процестерді қамтитын және болашақ нәтижелерді болжау үшін өткен деректер мен түсініктерді жалпылай алатын жасанды интеллект саласы. Негізінде, машиналық оқыту — бұл ақпарат алу, заңдылықтарды анықтау және деректерден қорытындылар жасау процесін қамтамасыз ететін компьютерлік жүйелерде жүзеге асырылатын математикалық әдістердің жиынтығы. Мақалада нақты бір зиянды бағдарламалық жасақтаманың түріне талдау жасалынды және таңдап алынған белгілерге негізделіп машиналық оқыту алгоритмінің көмегімен жіктеледі. Мұнда оқыту үшін алынған белгілерді таңдаудың рөлі өте маңызды. Кіріс деректері сапалы болған кезде машиналық оқыту жақсы жұмыс істейді.

Түйін сөздер: ақпараттық қауіпсіздік, зиянды бағдарламаларды талдау, машиналық оқыту, белгілерді таңдау, kNN

© Г. Алпысбай[1*], А. Бедельбаев[1], О. Усатова[1,2], А. Жумабекова[1], Эдзард Хофиг[3], 2023

[1]КазНУ имени аль-Фараби, Алматы, Казахстан;

[2]Институт информационно-вычислительных технологий, Алматы, Казахстан;

[3]Берлинский университет прикладных наук и технологий, Берлин, Германия.

E-mail: gulbanu.alpysbay@gmail.com

ПРИМЕНЕНИЕ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ВРЕДОНОСНОГО ПО

Алпысбай Г.Е. — Преподаватель. КазНУ им. аль-Фараби. Факультет Информационных технологий. 050040. Алматы, Казахстан

E-mail: gulbanu.alpysbay@gmail.com. ORCID: 0000-0002-3766-2396;

Бедельбаев А.А. — Кандидат физико-математических наук, профессор, КазНУ им. аль-Фараби. Факультет Информационных технологий. 050040. Алматы, Казахстан

E-mail: agyn08@yandex.ru. ORCID: 0000-0001-9839-4156;

Усатова О.А. — PhD, главный ученый секретарь, старший научный сотрудник. Институт Информационно-вычислительных технологий. 050010. Алматы, Казахстан. И.о. доцента.

КазНУ им. аль-Фараби. Факультет Информационных технологий. 050040. Алматы, Казахстан
E-mail: uoa_olga@mail.ru. ORCID: 0000-0002-5276-6118;
Жумабекова А.Т. — Преподаватель, КазНУ им. аль-Фараби. Факультет Информационных технологий. 050040. Алматы, Казахстан
E-mail: zhumabekova2702@gmail.com. ОРЦИД: 0000-0003-4242-7988;
Эдзард Хофиг — Профессор, доктор технических наук, Берлинский университет прикладных наук и технологий. Департамент медиа и компьютерных наук. Берлин, Германия
E-mail: edzard.hoefig@bht-berlin.de.

Аннотация. Быстрое развитие и широкое использование информационных технологий и Интернета имеют множество преимуществ, важно не игнорировать возникающие там негативные ситуации. Одной из таких негативных ситуаций является распространение и развитие вредоносных программ, приводящих к выходу из строя многих устройств и программного оборудования, разглашению конфиденциальной информации. Вредоносное ПО является средством для многих компьютерных атак и нарушений безопасности. Анализ вредоносных программ использует методы из нескольких различных областей, таких как анализ программ и сетевой анализ для изучения вредоносных шаблонов, чтобы получить более глубокое понимание нескольких аспектов, включая их поведение и то, как они развиваются с течением времени. Ежедневное развитие вредоносных программных средств, увеличение типов и сложности их структуры значительно усложнили работу по их выявлению и классификации. По этой причине возрастает важность использования систем искусственного интеллекта, в том числе алгоритмов машинного обучения, в сфере защиты информации. Машинное обучение — это область искусственного интеллекта, которая включает алгоритмы и процессы, которые «обучаются» и могут обобщать прошлые данные и идеи для прогнозирования будущих результатов. По сути, машинное обучение — это набор математических методов, реализованных в компьютерных системах, которые обеспечивают процесс получения информации, выявления закономерностей и вывода из данных. В статье анализируется тип вредоносного ПО в определенном формате и классифицируется с помощью алгоритма машинного обучения на основе конкретных выбранных признаков. Здесь очень важна роль отбора полученных признаков для обучения. Машинное обучение работает хорошо, когда входные данные хороши.

Ключевые слова: информационная безопасность, анализ вредоносных программ, машинное обучение, выбор признаков, kNN

Introduction

Malware is program that harms a user, computer or network and includes viruses, trojans, worms, rootkits, spyware, adware, etc. Studying the behavior of malicious programs remains an urgent problem in the field of information security. Since the appearance of the first malware, the variety of malware, its complexity, the number of new models and the speed with which they appear, as well as the range

of threats, have increased. The number of malicious programs in existing databases is huge, and new malicious programs appear every day. For example, the German laboratory AV-TEST registers more than 450,000 new malicious programs per day. Traditional anti-virus (reactive protection) based on checking signatures against a database of malicious codes cannot cope with emerging, previously unknown threats supplemented by an expert analyst. Therefore, tools for static analysis of a portable executable (PE) file and dynamic analysis of malware are emerging that can study pattern behavior that is not stored in a signature database. In addition, methods of working with anti-virus scanners by malicious programs are becoming more complicated (AV-TEST, 2022).

Malware analysis is the process of taking programs apart to understand how they work, and how to detect, disable, and remove them. There are millions of malwares on the Internet and this number is growing every day, so analyzing them is very important for everyone responsible for computer security. There are two main approaches to malware analysis: static and dynamic. Static analysis is the study of malware without running it. During dynamic analysis, the malware must be running. Both categories include basic and advanced techniques (Egele et al., 2012). The malware analysis process can often be sped up by making an educated guess about the malware's target and then confirming it. Of course, the accuracy of your predictions will depend on your knowledge of what malware typically does.

Methods and materials

Machine learning is a class of artificial intelligence methods, the characteristic feature of which is not the direct solution of a problem but learning by applying solutions to many similar problems. From the point of view of the application of machine learning in the field of information security, it can be classified into two main categories: pattern recognition and anomaly detection. It is not always possible to distinguish between pattern recognition and anomaly detection, but there is a well-defined task for each specific problem. In pattern recognition, we try to find obvious or undefined characteristics hidden in the data. These characteristics, labeled and combined into a set of features, can be used to train an algorithm that identifies the type of data with the same set of characteristics. Anomaly detection is the exact opposite of the task of obtaining data (Clarence Chio et al., 2018; Michael Sikorski et al., 2018). The main goal here is to determine the concept of normality that describes most of the data in the studied set instead of learning the characteristic patterns in the actual data set. After that, a deviation from the established normality is considered an anomaly.

A common misconception is that anomaly detection is the process of recognizing a set of "normal" patterns and distinguishing them from a set of "abnormal" patterns. The patterns obtained by the pattern recognition method must, of course, be taken from the studied data used to pre-train (train) the algorithm. On the other hand, when using the anomaly detection methodology, there may be an infinite number of anomalous samples with characteristics that match the given description of the

errors (outliers), even from hypothetical data that do not actually exist in the study (Ucci et al., 2017).

Malicious code detection by machine learning consists of several stages (figure 1):

— Obtaining information from the operating system about the files, processes, and software in use. All the data required for this machine learning can be obtained using Process Monitor and Process Hacker utilities for Windows operating system.

— Preparation of data for machine learning.

— Normalization of data received during work with files, registry, deletion of unnecessary settings (feature selection).

— Using machine learning methods.

— Output of results. Issuing the results of determining whether the process is harmful or not harmful.

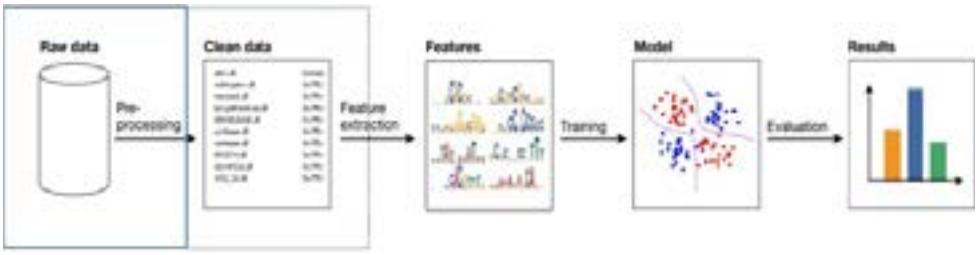

Fig. 1. Stages of detecting malicious code using machine learning

Malware can inject itself into different binary file formats that work in completely different ways. For example, PE files (extensions exe., dll., efi, i.e. portable executable files) in the Windows operating system, which we will discuss in our experiments throughout this article, have completely different internal file structures and require different execution contexts. Of course, the specific additional requirements for parsing each class of executables also differ significantly. Also note that malware can come in forms other than standalone executable binaries. There are common malicious components that infiltrate doc, pdf and rtf document files and use macros and dynamic executables in the document structure to perform malicious actions. Malware can also take the form of extensions and plug-ins for common software platforms, such as web browsers and complex web environments (Dilhara, 2021; Damin Moon et al., 2021).

The reverse engineering method is used to use the data with malicious program codes during machine learning, and since it is a large and extensive work, we preferred to work with a ready-made database. Open access malware databases:

- MalwareTrafficAnalysis.net website contains comprehensive, fully researched 600 samples of malicious code.

- VirusShare.com site 30 million provides an integrated database of malicious code patterns;

- VX Heaven team collected 270,000 malicious code samples for scientific use;
- In 2015, Kaggle and Microsoft managed to collect more than 10,000 malicious code samples into a single database by organizing the Malware Classification Challenge project.

In most cases, mindlessly loading a huge number of features into machine learning algorithms creates unnecessary noise and detrimentally affects the accuracy and efficiency of the model. Therefore, it is important to select only the most important and significant features for use in learning algorithms. This process is commonly known as feature selection. Feature selection can be done manually, based on domain expertise and information gained from the data mining phase. Features can be selected automatically using statistical methods and algorithms. In addition, there are unsupervised machine learning techniques often used for deep learning. (Malware Traffic Analysis, 2022).

One of the widely used methods of feature selection is the use of human practical experience. Human experts can provide a process for guiding machine learning models, which manifests itself primarily in the form of manually mined features, which are considered the most important pieces of information used in the human learning process. (VirusShare, 2022).

Classification of features depending on a specific model (model-specific feature ranking): when machine learning algorithms are applied to a certain set of data, the model of the final assessment can sometimes be expressed in the form of symbolic combinations of input features. (VX Heaven, 2010). For example, for a linear regression model, in which the value of Y is predicted based on a three-dimensional data set (in which the features are designated as , , and), the regression model can be represented by the following formula (without accounting for deviations):

$$Y = W_a x_a + W_b x_b + W_c x_c \text{Y} \quad (1)$$

After the training phase, the coefficients (weights) and will be assigned some values, for example:

$$Y = 3{,}72x_a + 1{,}94x_b + 0{,}138x_c \quad (2)$$

Even in this simplest example, it can be clearly seen that features and have higher weights than feature . Assuming that these features are appropriately normalized (their values are comparable values), you can eliminate the feature , knowing that it will not significantly affect the performance of the regression model.

Table 1. Features

№	Feature name	№	Feature name
1	sha256	26	minor_operating_system_version name
2	appeared	27	major_subsystem_version

3	label	28	minor_subsystem_version
4	file_size	29	sizeof_code
5	vsize	30	sizeof_headers
6	has_debug	31	sizeof_heap_commit
7	exports	32	imports
8	imports	33	exports
9	has_relocations	34	entry
10	has_resources	35	name_of_section
11	has_signature	36	size_of_section
12	has_tls	37	vsize_of_section
13	symbols	38	entropy
14	header	39	props
15	timestamp	40	histogram
16	machine	41	byte_entropy
17	characteristics	42	strings
18	subsystem	43	num_strings
19	dll_characteristics	44	avlength
20	magic	45	printabledist
21	major_image_version	46	printables
22	minor_image_version	47	paths
23	major_linker_version	48	urls
24	minor_linker_version	49	registry
25	major_operating_system_version	50	MZ

For example, let's say we have the following symbols describing a file in the PE format:

– histogram of byte values – a histogram of the distribution of byte values in a binary file;

– byte entropy histogram – a two-dimensional histogram of byte entropy approximating “the combined probability of containing a value in a byte and local entropy”;

– strings: an array of string statistics extracted from the original byte stream, defined by five or more consecutive characters with ASCII values between 0x20 (space) and 0x7f (del), or special strings such as C:\, HKEY_, http:// , and contains characteristics such as number of lines, average line length, number of C:\ paths, URL instances, registry keys, and a character distribution histogram;

– some “syntactically parsed characteristics”, for example:

• general information about the file - high-level details about the PE file, such as whether it was compiled with debug symbols, number of functions exported/ imported, etc.;

• file header information - details that can be obtained from the header section of a PE file related to the computer, architecture, OS, linker, etc.;

• information about sections of a binary file - names, sizes, entropy;

• import information – information about imported libraries and functions that can be used in the analyzed PE file;

• export information – information about functions exported from the analyzed PE file.

From among the above-mentioned signs, we further sort out the signs we need and prepare the data necessary for machine learning.

We will use these sets of extracted features (table 1) to classify files into malicious or benign types. To do this, we use the k-nearest neighbor algorithm.

The k-nearest neighbors (kNN) algorithm is the most widely known example of a lazy learning technique. Due to its simplicity, the kNN method is often used as a practical example to introduce machine learning concepts. This type of machine learning technique defers most of the computation to classification time instead of doing the work during training. (Kaggle and Microsoft, 2018). Lazy learning models do not learn data generalizations during the training phase. Instead, they capture (record) all the training data points that are passed to them and use this information to create local generalizations on the test set during classification. (Benjamin Bengfort et al., 2019).

The k nearest neighbors method is one of the simplest machine learning algorithms with the following characteristics: (Hyrum et al., 2018):

- the training stage consists of a simple recording (fixation) of all feature vectors and corresponding label elements in this model;

- the classification prediction is simply the most frequently occurring label among the k nearest neighbors in the test sample (according to the name of the technique).

Distance metrics for determining how “close” points are to each other in an n-dimensional feature space (where n is the size of the feature vectors) are typically the Euclidean metric (distance) for continuous variables and the Hamming distance for discrete variables. (Henrik Brink et al., 2017).

Results

The classification report is presented in Table 2.

Table 2. Classification report

	Precision	Recall	F1-score	Support
0	0.90	0.78	0.83	7458
1	0.84	0.94	0.89	9711
avg/total	0.87	0.87	0.86	17169

Accuracy is our most common evaluation metric and is easy to understand, i.e. the number of samples to be matched divided by the number of all samples. In general, the higher the accuracy, the better the classifier. The degree of accuracy is indeed a very good and intuitive measure of the estimate, but sometimes a high degree of accuracy does not reflect the algorithm. The accuracy of the model was ***0.866620071058***.

Precision is for the results of our predictions and shows how many samples whose predictions are positive are correct. Then there are two possibilities to predict the positive class, one is to predict the positive class as class positive (TP) and the other is to predict the negative class as class positive (FP).

Recall is for our original sample and indicates how many positive examples in the sample are predicted correctly. There are also two possibilities, one is to predict the original positive class as class positive (TP) and the other is to predict the original positive class as class negative (FN). The recall rate is a measure of coverage.

F1-score. Indicators Precision and Recall sometimes have contradictions, so they need to be considered comprehensively. The most common method is F-Measure (also known as F-Score). F-Measure is the weighted harmonic mean of precision and recall.

Support is the number of instances of each class.

Conclusion

This article provides information on how high the level of distribution of malicious software equipment is at the present time and the importance of their detection using machine learning algorithms, as well as data obtained from practical work. The classifier model is trained on data normalized by feature importance. The trained model takes only external observation data (on the end user's machine) as input. Based on these data, the model tries to reconstruct the complete pattern of events. Model training and training is an independent study. The machine learning module is constantly trained on the results of running the suspicious software, and the model on end-user machines is updated in a continuous integration mode.

REFERENCES

AV-TEST: The Independent IT-Security Institute, 2022. Malware. AV-TEST URL: https://www. av-test.org/en/statistics/malware/

Egele M., Scholte Th., Kirda E., Kruegel C., 2012. A survey on automated dynamic malware analysis techniques and tools. ACM Computing Surveys, 44(2). Pp. 22–28.

Wikipedia, 2022. Antivirus software. URL: https://en.wikipedia.org/wiki/Antivirus_software

Michael Sikorski, Andrew Honig, 2018. Practical Malware Analysis. The Hands-On Guide to Dissecting Malicious Software. San Francisco, DC: No Starch Press.

Clarence Chio, David Freeman, 2018. Machine Learning and Security: Protecting Systems with Data and Algorithms. Sebastopol, DC: O'Reilly.

Ucci D., Aniello L., Baldoni R., 2017. Survey on the Usage of Machine Learning Techniques for Malware Analysis. ACM Transactions on the Web, 1. Pp, 3–58.

B.A.S. Dilhara, 2021. Classification o f M alware u sing M achine l earning a nd D eep learning Techniques International Journal of Computer Applications (0975–8887). 183–32.

Damin Moon, JaeKoo Lee, MyungKeun Yoon, 2021. Compact feature hashing for machine learning based malware detection Information & Communications Technology Express, DOI: 10.1016/j.icte.2021.08.005.

Malware Traffic Analysis, 2022. A source for packet capture (pcap) files and malware samples. URL: https://www.malware-traffic-analysis.net/

VirusShare, 2022. URL: https://virusshare.com/

VX Heaven, 2010. VX Heaven Virus Collection. URL: https://web.archive.org/web/20170611163424/http://vxheaven.org/

Kaggle and Microsoft, 2018. Microsoft Malware Prediction. URL: https://www.kaggle.com/c/microsoft-malware-prediction

Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda, 2019. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. Sebastopol, DC: O'Reilly.

Hyrum S. Anderson, Phil Roth, 2018. An Open Dataset for Training Static PE MalwareMachine Learning Models.

Henrik Brink, Joseph W. Richards, Mark Fetherolf, 2017. Real-World Machine Learning. Manning: Shelter, Island.

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Заместитель директор отдела издания научных журналов НАН РК *Р. Жәлиқызы*
Редакторы: *М.С. Ахметова, Д.С. Аленов*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 12.06.2023.
Формат 60х881/8. Бумага офсетная. Печать –ризограф.
19,0 п.л. Тираж 300. Заказ 2.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19