Автоматика. Информатика. Управление. Приборы

УДК 004.42
DOI: 10.17277/vestnik.2018.01.pp.006-015

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ЭВРИСТИЧЕСКОГО АНАЛИЗА В АНТИВИРУСНОМ ПАКЕТЕ STRONGHOLD ANTIMALWARE

Р. Ю. Демина, И. М. Ажмухамедов

Кафедра «Информационная безопасность»,
ФГБОУ ВО «Астраханский государственный университет»,
г. Астрахань, Россия;
raisapereverzeva@gmail.com

Ключевые слова: антивирусный анализ; бинарная классификация; обучающее множество; эвристический анализ.

Аннотация: Рассмотрены результаты внедрения специально разработанной методики формирования обучающего множества в деятельность антивирусной компании Security Stronghold LLC. Показано увеличение эффективности блока эвристического анализа в 2 раза. Рекомендовано применение упомянутой методики для обучения эвристических классификаторов других антивирусных пакетов.

Введение

Одними из основных инструментов защиты пользовательских данных, хранящихся на ПЭВМ, являются антивирусные пакеты. Установленные на «машине» пользователя, они оказывают противодействие вредоносному коду, передающемуся по сети или через съемные носители.

Как правило, современные антивирусные пакеты предусматривают два основных типа сканирования: сигнатурный и эвристический анализы. *Сигнатурный анализ* является основным средством борьбы с вредоносными объектами и однозначно выявляет известные вирусы. Он предусматривает следующие этапы: идентификацию экспертами антивирусной компании неизвестной программы как вирусной; определение всех ее отличительных признаков; выявление модификаций, вносимых данной программой в систему; занесение информации о вирусе в базу данных сигнатур. Затем база данных «выкладывается» на сервер, и клиентские приложения запускают процесс обновления базы данных. Только после этого, пользователь будет защищен от предварительно выявленных экспертами угроз. Основным недостатком антивирусного анализа является невозможность обнаруживать вирусы «нулевого дня», то есть те угрозы, которые не проанализированы экспертами и чьи сигнатуры отсутствуют в базе данных.

Данного недостатка не имеет *эвристический анализ*, который с некоторой вероятностью может отнести сканируемый файл к категории вредоносных или

легитимных объектов. Выделяют два основных типа эвристического анализа: динамический и статический. В ходе *динамического* эвристического анализа сканируемый файл запускается в безопасном виртуальном пространстве, так называемой «песочнице», после чего антивирус анализирует его действия, производимые в операционной системе. Основной недостаток динамического анализа – его требовательность к вычислительным ресурсам для эмуляции операционной системы. Им не обладает *статический* эвристический анализ, в ходе которого рассматривается структура и содержимое файла и выявляются признаки, характерные для других, ранее изученных вирусов. В основе статического эвристического анализа лежит задача бинарной классификации, состоящая из двух основных этапов: обучения классификатора и распознавания (определения является ли неизвестный файл вредоносным или легитимным). Этап обучения – первоочередной и во многом предопределяет верность классификации (долю правильно распознанных объектов).

Несмотря на широкое применение статический эвристический анализ распознает не более 60 % неизвестных вирусов. Как свидетельствуют результаты ранее проведенных экспериментов [1] и имеющиеся в научно-технической литературе данные [2, 3], качество бинарной классификации (задачи лежащей в основе эвристического анализа) во многом зависит от состава обучающего множества (**ОМ**). Исходя из этого, в работе [4] предложена специальная методика формирования ОМ, которая может использоваться для повышения эффективности бинарной классификации, и, как следствие, антивирусного эвристического анализа. Для полноценной оценки ее эффективности необходима проверка данной методики в реальных условиях, то есть требуется обучить эвристический классификатор, который входит в состав антивирусного пакета, достаточно широко используемого пользователями. Таким образом, сформулирована задача, результаты решения которой изложены в данной статье.

Постановка задачи

Необходимо оценить повышение эффективности эвристического статического анализа при использовании на стадии обучения методики формирования ОМ, предложенной ранее в [4].

Решение задачи

Обучающее множество, используемое для обучения классификатора, как правило, формируется случайным образом из имеющихся в наличии маркированных объектов, то есть таких, для которых определено к какому классу они относятся. Отобранные объекты в таком случае становятся своего рода эталонами, на которые при распознавании будет ориентироваться классификатор.

В работе [4] для отбора наиболее подходящих для обучения эвристического статического классификатора файлов были предложены критерий для сравнения кандидатов в ОМ и алгоритм отбора файлов в обучающее множество.

Критерий сравнения файлов

Для целенаправленного формирования ОМ необходимо иметь возможность сравнивать между собой файлы. Критерием сравнения выбрана байтовая структура файлов [4].

В качестве меры схожести двух файлов предложено использовать величину

$$\rho(A,B)=\frac{\left|\overline{A}\cap\overline{B}\right|}{\overline{A}},$$

где $\overline{A}$ – множество n-грамм файла A; $\overline{B}$ – множество n-грамм файла B.

Из данной формулы мера схожести файлов A и B определяется как отношение числа уникальных n-грамм (любых рядом стоящих n байт) файла A, которые встречаются в наборе n-грамм файла B, к числу уникальных n-грамм файла A.

Введенная таким образом мера позволяет построить матрицу схожести для множества файлов F: квадратную матрицу, состоящую из элементов ρ_{ij} – мер схожести i-го файла с j-м. При этом данная матрица в общем случае асимметрична, так как $\rho(A, B)$ может быть неравен $\rho(B, A)$.

Алгоритм отбора файлов в обучающем множестве

Матрица схожести служит основой алгоритма отбора максимально различающихся между собой по составу n-грамм файлов во множество F' для использования на этапе обучения классификаторов ($|F| \geq |F'|$). Для этого необходимо:

1. Исходя из допустимого времени обучения классификатора, целей обучения и алгоритма классификации, задать M – мощность множества F'.

2. Задать пороговое значение меры схожести K, при превышении которого сравниваемые файлы считаются схожими.

3. Для каждого i-го файла подсчитать параметр L_i – число j-х файлов ($i \neq j$), для которых $\rho_{ij} \geq K$, тем самым выделить группы взаимно схожих файлов. Файлы с $L_i = 0$ считаются уникальными в рамках данного множества.

4. Из каждой группы произвольно выбрать и оставить только один файл.

5. Включать файлы, отобранные на шаге 4, в итоговую обучающую выборку F' в порядке убывания параметра L до тех пор, пока $|F'| \leq M$.

6. Если после выполнения шага 5 $|F'| \leq M$, то необходимо понизить пороговое значение K и повторить шаги 2 – 5. В противном случае – множество F' считается сформированным, и работа алгоритма заканчивается.

Таким образом, методика отбора представляет собой механизм кластеризации, то есть обучение без учителя. Она позволяет разделить исходный набор вредоносных файлов на взаимно сходные классы по определенным признакам в отсутствие информации об используемых ими механизмах вредоносного воздействия.

Применение методики формирования ОМ позволяет повысить верность распознавания антивирусного эвристического классификатора, но при этом увеличивает общее время этапа обучения. Для сокращения времени расчета матрицы схожести (наиболее затратной операции методики формирования обучающего множества) целесообразно учитывать ряд особенностей программной реализации [5]: предварительное составление перечня уникальных n-грамм, одновременный расчет $\rho(A, B)$ и $\rho(B, A)$, использование многопоточности. С учетом представленных рекомендаций разработано соответствующее программное обеспечение [6], позволяющее целенаправленно формировать ОМ за минимальное время. Для оценки целесообразности внедрения разработанной методики в деятельность антивирусной компании проведена серия экспериментов в лабораторных условиях.

Проверка эффективности применения предложенной методики в лабораторных условиях

Для проверки эффективности методики проведена серия вычислительных экспериментов. Для формирования ОМ использовался специально разработанный в среде Microsoft Visual Studio 2015 программный продукт на языке C++.

Для построения классификатора и последующего распознавания использовался алгоритм AdaBoost (*Adaptive Boosting*) [7, 8] в рамках программного комплекса Weka 3.8 [9].

Обучение проведено на двух наборах файлов, при этом множество легитимных файлов в них одинаковое и составляло 1000 шт. Кроме легитимных, в первый набор (Tr_{method}) вошли вредоносные файлы, специально отобранные по предложенной выше методике (1000 шт.). Второй набор (Tr_{rand}) включал в себя случайно отобранные вирусные файлы (1000 шт.). Отбор вредоносных файлов для обоих множеств проводился из вирусного набора, содержащего в себе 3000 файлов. Для тестирования всего отобрано 5000 легитимных и 5000 вредоносных файлов, из которых сформировано случайным образом по пять соответствующих тестовых наборов, имеющих 2 тысячи объектов в каждом. При проведении экспериментов эффективность классификатора оценивалась числом правильно классифицированных вредоносных файлов *TN* и числом правильно классифицированных легитимных файлов *TP*. В таблице 1 представлены сравнительные результаты распознавания.

Очевидно, что применение предложенной методики привело к стабильному увеличению параметра *TN* (повысилось обнаружение вирусов) в среднем на 9,49 % во всех контрольных группах. При этом, как и ожидалось, увеличение *TN* повлекло за собой незначительное (на 0,36 %) уменьшение *TP* (увеличилось число ложных срабатываний на легитимные файлы).

В таблицах 2 и 3 показаны сопряженности для третьего тестового набора (как наиболее соответствующего усредненным показателям) для случаев с обучением на Tr_{rand} и Tr_{method}. Каждая строка данных таблиц соответствует фактическим классам из проверочного множества, каждый столбец – классам, предсказанным классификатором. В последнем столбце и последней строке находятся так называемые «маргиналы» [3] – суммы элементов в соответствующем столбце или строке, значения которых позволяют оценить статистическую значимость полученных результатов.

Таблица 1

Результаты распознавания для различных наборов, %

Номер набора	*TN*					*TP*				
	1	2	3	4	5	1	2	3	4	5
Tr_{method}	79,45	79,2	78,6	78,95	79,55	97,6	96,05	98,7	98	97,75
Tr_{rand}	69,15	68,2	69,45	70,15	71,35	98,15	96,25	99,05	98,45	98
Разница показателей	10,30	11,0	9,15	8,80	8,20	–0,55	–0,20	–0,35	–0,45	–0,25

Таблица 2

Таблица сопряженности для Tr_{rand}

Проверочное множество	Предсказано «легитимный»	Предсказано «вирус»	Маргиналы
Фактически «легитимный»	1981	19	2000
Фактически «вирус»	611	1389	
Маргиналы	2592	1408	4000

Таблица 3

Таблица сопряженности для Tr_{method}

Проверочное множество	Предсказано «легитимный»	Предсказано «вирус»	Маргиналы
Фактически «легитимный»	1974	26	2000
Фактически «вирус»	428	1572	
Маргиналы	2402	1598	4000

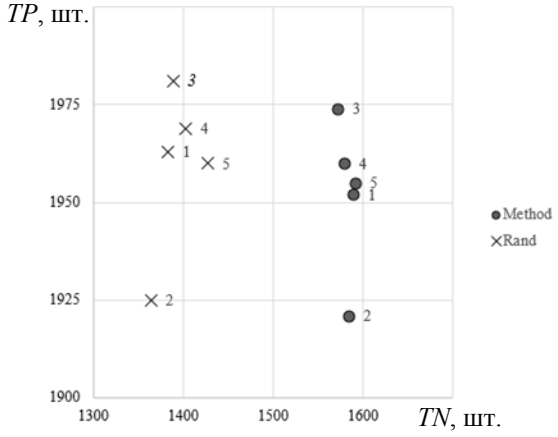

Рис. 1. График покрытия

На рисунке 1 показан график покрытия, на котором представлены результаты распознавания для двух групп классификаторов: Rand, обученных на Tr_{rand}, и Method, обученных на Tr_{method}.

Несмотря на то что Method совершает на семь ошибок больше при обнаружении легитимных файлов, он превосходит Rand по распознаванию вирусов на 183 шт.

Еще одной характеристикой качества классификатора является верность – доля правильно классифицированных тестовых примеров:

$$\text{acc} = \frac{1}{|Te|}\sum_{x \in Te} I\left[\hat{c}(x) = c(x)\right],$$

где $|Te|$ – мощность проверочного множества; I – индикаторная функция, которая равна 1, если оценочная метка класса $\hat{c}(x)$ совпадает с истинной меткой класса $c(x)$, то есть если объект был правильно классифицирован.

Например, для данных из табл. 2: acc = 0,8425 или 84,25 %; табл. 3: acc = 0,8865 или 88,65 %.

Также рассчитана частота ошибок – доля неправильно классифицированных примеров

$$\text{err} = \frac{1}{|Te|}\sum_{x \in Te} I\left[\hat{c}(x) \neq c(x)\right].$$

Для данных из табл. 2: err = 0,1575 или 15,75 %; табл. 3: err = 0,1135 или 11,35 %.

Таким образом, несмотря на то что классификатор Method не доминирует над Rand (не превосходит его по всем показателям), его верность на 4,4 % больше.

Проверка эффективности методики формирования обучающего множества в реальных («боевых») условиях

Проверка рассматриваемой методики в реальных условиях осуществлена в рамках антивирусного программного продукта Stronghold AntiMalware [10], который является наиболее динамично развивающимся продуктом компании Security Stronghold. В своем составе Stronghold AntiMalware содержит следующие основные блоки: сигнатурный анализ, активный монитор, модули удаления вредоносных объектов и восстановление пользовательских настроек системы, а также блок статического эвристического анализа.

Эффективность данного модуля оценивалась следующим образом. За вторую половину 2015 г. собрана статистика о результатах 11 690 пользовательских сканирований. В среднем за каждое сканирование проанализировано 31 349 файлов, из которых сигнатурный анализ выявлял по 28 вирусных файлов. Папки, ключи и значения реестра, а также поврежденные файлы конфигураций не учитывались. Эвристика из файлов, признанных сигнатурным анализом легитимными, дополнительно относила к потенциально опасным еще по четыре файла. Последующий анализ данных файлов, более подробная информация о которых также отправлялась на сервер компании со статистикой, показал, что в среднем один файл был отнесен к вредоносным по ошибке и на самом деле являлся легитимным. В ходе сбора статистики на сервер отправлялась следующая информация о каждом файле, признанным модулем эвристического сканирования вредоносным: хеш-сумма файла, его полное имя, сертификат и пр.

Решение о том, ошибся ли эвристический анализ или нет, принималось следующим образом: через два месяца после окончания сбора статистики собранные значения хеш-сумм загружались для проверки с помощью сервиса Virus Total [11]. Данный сервис позволяет получить информацию о том, какие антивирусы отнесли файл с такой же хеш-суммой к классу вредоносных объектов. При этом добавление к сигнатурному анализу модуля эвристики увеличило среднее время сканирования с 20 минут до 25.

Таким образом, оценка эффективности блока эвристики за 2015 г. показала, что количество выявленных у реальных пользователей вирусов «нулевого дня» слишком мало и не стоит затраченного на это времени сканирования.

Исходя из полученных результатов исследования в реальных и лабораторных условиях, принято решение о необходимости увеличить число корректно распознанных вирусов «нулевого дня» с помощью вышепредствленной методики формирования ОМ. После внедрения методики формирования ОМ для обучения антивирусного классификатора пакета Stronghold AntiMalware во второй половине 2016 г. собрана статистика, аналогичная составленной во второй половине 2015 г. Сравнительные результаты представлены в табл. 4.

Кроме того, в целях сокращения времени эвристического сканирования в процесс формирования перечня признаков сканируемого файла внедрен расширенный алгоритм бинарного поиска [12]. Среднее время сканирования в результате сократилось с 25 минут до 23 (на 8 %).

Таблица 4

Результаты применения методики формирования обучающего множества в компании Sequirity Stronghold (июнь – декабрь 2015 г.)

Результаты пользовательских сканирований	Случайное формирование ОМ	Формирование ОМ с помощью методики «Формирование ОМ для задач статического эвристического анализа»
Число сканирований	11690	14326
Усредненные показатели результатов сканирований		
Просканировано файлов	31349	31490
Обнаружение вирусных файлов: – сигнатурный анализ	28	31
– эвристический анализ: всего:	4	7
правильно	3	6
неправильно	1	1

Обсуждение результатов

Из таблицы 4 очевидно, что внедрение методики формирования позволило увеличить число корректно распознанных вирусов «нулевого дня» в 2 раза. Поскольку время, затраченное на эвристическое сканирование за счет внедрения расширенного алгоритма бинарного поиска, сократилось на 8 %, можно утверждать, что полученные данные характеризуют степень повышения эффективности блока эвристического сканирования за счет применения методики целенаправленного отбора файлов в ОМ.

Выводы

Таким образом, для повышения эффективности эвристического антивирусного анализа, который является важной составляющей антивирусных пакетов, целесообразно использовать предложенную ранее методику отбора файлов в обучающее множество. Экспериментальная проверка применения данной методики на примере Stronghold AntiMalware показала увеличение эффективности блока эвристического анализа в 2 раза, что позволяет рекомендовать применение упомянутой методики для обучения эвристических классификаторов других антивирусных пакетов.

Список литературы

1. Демина, Р. Ю. Зависимость эффективности обнаружения вредоносного программного обеспечения от качества обучающей выборки в алгоритмах классификации / Р. Ю. Демина, И. М. Ажмухамедов // Математические методы в технике и технологиях – ММТТ-28 : сб. тр. XXVIII Междунар. науч. конф. : в 12 т. – Саратов : СГТУ, 2015. – Т. 3. – С. 64 – 66.

2. Потапов, А. С. Распознавание образов и машинное восприятие / А. С. Потапов. – СПб. : Политехника, 2007. – 552 с.

3. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – М. : ДМК Пресс, 2015. – 400 с.

4. Демина, Р. Ю. Методика формирования обучающего множества при использовании статических антивирусных методов эвристического анализа / Р. Ю. Демина, И. М. Ажмухамедов // Инженерный вестник Дона. – 2015. – № 3. – Режим доступа : http://ivdon.ru/uploads/article/pdf/IVD_204_demina_azhmuhamedov.pdf_0b8ea4a2fc.pdf (дата обращения: 16.03.2015).

5. Демина, Р. Ю. Особенности программной реализации алгоритмов методики формирования обучающего множества для бинарных классификаторов, используемых в антивирусном эвристическом статическом анализе / Р. Ю. Демина // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2017. – № 2. – С. 62 – 68.

6. Свидетельство о гос. рег. программ для ЭВМ № 2015662690. Формирование обучающего множества для задач статического эвристического анализа / Р. Ю. Демина, И. М. Ажмухамедов. Зарегистр. в реестре программ для ЭВМ. – 13 октября 2015 г.

7. Harrington, P. Machine Learning in Action / P. Harrington. – Manning Publications Co, 2012. – 382 p.

8. Marsland, S. Machine Learning: An Algorithmic Perspective / S. Marsland. – Chapman and Hall/CRC, 2014. – 457 p.

9. Weka 3: Data Mining Software in Java. – Режим доступа : http://www.cs.waikato.ac.nz/ml/weka/ (дата обращения: 14.06.2017).

10. Сканер-антишпион Stronghold AntiMalware // Security Stronghold. – Режим доступа : https://www.securitystronghold.com/ru/stronghold-antimalware/ (дата обращения: 14.06.2017).

11. Virustotal. – Режим доступа : https://www.virustotal.com/ (дата обращения: 14.06.2017).

12. Демина, Р. Ю. Метод сокращения времени обучения антивирусного эвристического классификатора, основанный на использовании алгоритма расширенного бинарного поиска / Р. Ю. Демина, И. М. Ажмухамедов, Т. Г. Гурская // Прикаспийский журнал: управление и высокие технологии. – 2017. – № 1. – С. 15 – 23.

Improving the Efficiency of Heuristic Analysis in the Anti-Virus Stronghold Antimalware

R. Yu. Demina, I. M. Azhmukhamedov

Department of Information Security,
Astrakhan State University, Astrakhan, Russia;
raisapereverzeva@gmail.com

Keywords: antimalware; heuristic; training set; binary classification.

Abstract: The heuristic analysis is one of the most important parts of modern anti-virus packages. It can detect viruses of “zero day”. Binary classification is the cornerstone of the heuristic analysis. Training and detecting are two main stages of the classification. At the same time, the training set influences the result of classification. The article describes the technique of formation of a training set composed of different files. The effectiveness of proposed techniques was assessed under laboratory conditions. The experiments revealed the viability of implementing the techniques in the antimalware laboratory SecurityStronghold. The technique of formation of a training set for learning anti-virus heuristic classification increased the number of true detected malwares by two times.

References

1. Demina R.Yu., Azhmukhamedov I.M. [Dependence of detection efficiency of malicious software on the quality of training sample in classification algorithms], *Matematicheskie metody v tekhnike i tekhnologiyakh – MMTT-28* [Mathematical Methods in Engineering and Technology - MMTT-28], vol. 3. Saratov: SGTU, 2015, pp. 64-66 (In Russ.)
2. Potapov A.S. *Raspoznavanie obrazov i mashinnoe vospriyatie* [Image recognition and machine perception], St. Petersburg: Politekhnika, 2007, 552 p. (In Russ.)
3. Flakh P. *Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh* [Machine learning. The science and art of constructing algorithms that extract knowledge from data], Moscow: DMK Press, 2015. 400 p. (In Russ.)
4. Demina R.Yu., Azhmukhamedov I.M. [The method of forming a training set using static antivirus methods of heuristic analysis], *Inzhenernyi vestn. Dona* [The engineer's messenger of the Don], 2015, no. 3, available at: http://ivdon.ru/uploads/article/pdf/IVD_204_demina_azhmuhamedov.pdf_0b8ea4a2fc.pdf (accessed 16 Mart 2017).
5. Demina R.Yu. [Features of the software implementation of algorithms for the formation of training set for binary classifiers used in antivirus heuristic static analysis] *Vestnik AGTU* [Bulletin of the Astakhan State Technical University], 2017, no. 2, pp. 62-68 (In Russ.)
6. Demina R.Yu., Azhmukhamedov I.M. *Formirovanie obuchayushchego mnozhestva dlya zadach staticheskogo evristicheskogo analiza* [Formation of training set for tasks of static heuristic analysis], Russian Federation, 2015, Certificate of state registration of computer programs No. 2015662690 (In Russ.)
7. Harrington P. *Machine Learning in Action*, Manning Publications Co, 2012, 382 p.
8. Marsland S. *Machine Learning: An Algorithmic Perspective,* Chapman and Hall/CRC, 2014, 457 p.
9. http://www.cs.waikato.ac.nz/ml/weka/ (accessed 14 June 2017).
10. https://www.securitystronghold.com/ru/stronghold-antimalware/ (accessed 14 June 2017).
11. https://www.virustotal.com/ (accessed 14 June 2017).
12. Demina R.Yu., Azhmukhamedov I.M., Gurskaya T.G. [The method of reducing the training time of the anti-virus heuristic classifier, based on the use of the extended binary search algorithm], *Prikaspiiskii zhurnal: upravlenie i vysokie tekhnologii* [Pri-Caspian Journal: Management and High Technologies], 2017, no. 1, pp. 15-23 (In Russ.)

Erhöhung der Effizienz der heuristischen Analyse im Antiviruspaket Stronghold Antimalware

Zusammenfassung: Einer der wichtigsten Bestandteile moderner Antivirus-Pakete ist die heuristische Analyse, die "Zero Day" Viren erkennen kann. Die heuristische Analyse basiert auf dem Problem der binären Klassifikation. Dabei wird das Ergebnis der Klassifikation wesentlich durch die Zusammensetzung des Trainingssatzes beeinflusst. Es sind die Ergebnisse der Einführung der speziell entwickelten Methodik zur Bildung von Trainingssätzen in der Tätigkeit des Antiviren-Unternehmens Security Stronghold LLC betrachtet.

Augmentation de l'efficacité de l'analyse heuristique dans un paquet antivirus stronghold anti malware

Résumé: L'une des plus importantes parties des modernes paquets antivirus est l'analyse heuristique capable de reconnaître les virus de type «jour zéro». A la base de l'analyse heuristique se trouve une tâche de la classification binaire. La composition de l'ensemble d'apprentissage influence suffisamment sur le résultat de la classification. Sont examinés les résultats de la mise en œuvre de la méthodologie spécialement conçue de la formation de l'ensemble d'apprentissage dans les activités de la société d'antivirus Security Stronghold LLC.

Авторы: ***Демина Раиса Юрьевна*** – ассистент кафедры «Информационная безопасность», ФГБОУ ВО «Астраханский государственный университет», аспирант кафедры «Информационная безопасность», ФГБОУ ВО «Астраханский государственный технический университет»; ***Ажмухамедов Искандар Маратович*** – доктор технических наук, доцент, заведующий кафедрой «Информационная безопасность», ФГБОУ ВО «Астраханский государственный университет», г. Астрахань, Россия.

Рецензент: ***Попов Георгий Александрович*** – доктор технических наук, профессор, заведующий кафедрой «Информационная безопасность», ФГБОУ ВО «Астраханский государственный технический университет», г. Астрахань, Россия.