УДК 004.056.57 Вестник УрФО № 1(47) / 2023, с. 69–80

Гибилинда Р. В., Коллеров А. С., Фартушный А. В. DOI: 10.14529/secur230107

АВТОМАТИЗАЦИЯ ПРОЦЕССА МАРКИРОВКИ ПАКЕТОВ СЕТЕВОГО ТРАФИКА СЕРВИСОВ МГНОВЕННОГО ОБМЕНА СООБЩЕНИЯМИ ПРИ РАССЛЕДОВАНИИ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

В статье предлагается решение задачи автоматизации процесса маркировки пакетов сетевого трафика в контексте выявления утечек информации. Маркировка сетевых пакетов зашифрованного трафика осуществляется с использованием методов кластеризации и классификации, инициируемых с помощью сервисов мгновенного обмена сообщениями. В основе решения лежит принцип «минимальной длины описания» (minimum description length – MDL), предназначенный для расчета оптимального количества кластеров. Авторами предлагается метод, позволяющий автоматизировать процесс отбора и маркировки пакетов сетевого трафика для формирования «обучающей выборки», с учетом обоснованного выбора размерности признакового пространства в целях повышения скорости работы метода, а также алгоритмической оценки результатов кластеризации.

Ключевые слова: *инцидент информационной безопасности, кластеризация, трафик, мессенджер.*

AUTOMATION OF INSTANT MESSAGING SERVICES NETWORK TRAFFIC PACKETS MARKING PROCESS IN INFORMATION SECURITY INCIDENTS INVESTIGATION

The article proposes a solution to the task of automating network traffic packets marking process. This solution allows marking network packets of encrypted traffic using clustering and classification methods for detecting information leaks initiated through instant messaging services. The proposed solution is based on the «minimum description length» principle (MDL), designed to calculate the optimal number of clusters. Authors propose network traffic packets marking and selection process automatization method for «training set» creation based on justified selection of feature space dimension in order to increase method working speed and algorithmic evaluation of clustering results.

Keywords: *information security incident, clustering, traffic, messenger.*

Введение

Деятельность любой компании приводит к появлению конфиденциальной информации (коммерческая тайна, персональные данные клиентов, банковская тайна и др.), для которой необходимо обеспечивать защиту от утечки. Наиболее освещаемые в средствах массовой информации утечки указанной информации происходят из-за наличия нелояльных к компании сотрудников (примеры статей, по которым были привлечены к уголовной ответственности сотрудники различных компаний – ч. 2 ст.138, ч. 3 ст. 272, ч. 4 ст. 274.1 УК РФ). Одним из каналов утечки, которым могут воспользоваться нелояльные сотрудники, являются сервисы мгновенного обмена сообщениями (далее – мессенджеры), установленные или запускаемые на рабочих станциях или корпоративных мобильных телефонах. При этом устройства взаимодействуют/работают в корпоративной сети с возможностью передачи информации в сеть Интернет. Согласно результатам исследования [1], 90% опрошенных компаний используют мессенджеры для решения рабочих вопросов (52% российских компаний используют некорпоративные мессенджеры для коммуникаций между сотрудниками). При использовании некорпоративных мессенджеров отсутствует возможность контроля передачи информации по зашифрованному каналу. В итоге специалистам по информационной безопасности организаций (далее – специалист) затруднительно своевременно реагировать и расследовать возникающие инциденты, связанные с утечками.

Существующие системы защиты от утечек (data leakage prevention – DLP) имеют ограниченный функционал (в области анализа зашифрованного трафика) [2], в связи с этим специалисту приходится анализировать поток пакетов сетевого трафика (далее – СТ), исходящий с рабочих устройств сотрудников для определения несанкционированного использования мессенджеров. Следует учитывать, что пакеты СТ мессенджеров, в подавляющем большинстве случаев, передаются в зашифрованном виде. Процесс такого анализа включает применение алгоритмов классификации для определения типа используемого мессенджера [3]. Обучение классификатора предполагает отбор и маркировку пакетов

СТ для формирования «обучающей выборки». В условиях больших объемов передаваемых данных и появления новых типов мессенджеров процесс «ручной» маркировки пакетов СТ становится трудозатратным. В связи с этим возникает потребность в автоматизированном процессе отбора пакетов СТ и их маркировки, позволяющем формировать наборы данных для последующей классификации в любых признаковых пространствах.

Авторами предлагается метод, позволяющий автоматизировать процесс отбора и маркировки пакетов СТ для формирования «обучающей выборки».

1. Выбор признакового пространства для подготовки СТ к обработке алгоритмами классификации

В проведенном ранее исследовании [3] анализировался зашифрованный СТ и предложено использование методов классификации для решения задачи идентификации мессенджеров. Для проведения экспериментов применялся СТ мессенджеров (с указанием количества пакетов): WhatsApp (43100), Discord (45631) и Skype (68453) в суммарном объеме 157184 пакетов. В качестве шума для имитации фоновой активности СТ использовались пакеты почтового клиента и веббраузера в суммарном объеме 108576. Указанный трафик был записан с помощью личных мобильных устройств авторов посредством «зеркалирования» пакетов с личных беспроводных точек доступа.

Учитывая объемы циркулирущего СТ в ходе проведенных экспериментов, а также наличие шифрования данных, для идентификации мессенджеров методами классификации необходимо автоматизировать процесс формирования «обучающей выборки».

В [3] обосновано использование только части пакета зашифрованного СТ без служебной информации, содержащего полезные данные, (полезная нагрузка, payload) для его дальнейшего анализа. Согласно [3], полезная нагрузка представлена в виде вектора переменной длины T = (b_1, b_2, b_3, ... b_h), где b_i – значение i-ого байта пакета, bi ∈ [0,255], h = $h = \overline{1,1460}$ и ограничена максимальным размером блока данных одного пакета (maximum transmission unit – MTU – 1460 байт) [4]. Размерность формируемых векторов приведена к значению 1460 (недостающие элементы дополняются символом 256, который не встречается в стандартных пакетах). Результирующее количество признаков, используемых для анализа, совпадает с размерностью вектора T. Таким образом преобразованный вектор T = (b_1, b_2, b_3, ... b_{1460}), где b_i – i-ый количественный признак, $b_i \in [0,256]$. Из векторов, являющихся промаркированными (соотнесенными с некоторыми классами – мессенджерами) пакетами СТ, создается «обучающая выборка», представленная в виде множества $\mathbb{T} = \{T_n | n = \overline{1,N}\}$, где N – количество векторов.

2. Представление пакетов СТ в виде точек, распределенных в пространстве. Выбор алгоритмов кластеризации

Метод оценки результатов группировки пакетов СТ для формирования «обучающей выборки», предлагаемый авторами, основан на использовании кластерного анализа для элементов множества $\mathbb{T}$. Группировка в дальнейшем позволит производить маркировку (соотнесение с классами – мессенджерами) не каждого пакета в отдельности, а их набора (группы). Рассмотрим подробнее процесс группировки.

Для начала вектор T_n взаимно-однозначно отображается в точку T'_n = (b_1, b_2, b_3, ... b_h), которая является элементом признакового пространства $\mathcal{B}^{1460}$, где $b_h \in [0,256]$.

Отображение пакетов СТ в виде точек T'_n в признаковом пространстве $\mathcal{B}^{1460}$ позволяет применять кластерный анализ с использованием широко распространенных неиерархических алгоритмов DBSCAN [5] и OPTICS [6]. Выбор алгоритмов обоснован произвольной геометрической формой кластеров, образуемых точек T'_n. Так как размерность признакового пространства превышает 3, то отображение данных, в возможном для восприятия трехмерном пространстве, не представляется реализуемым без их предварительной обработки, например, методом главных компонент (МГК, PCA – principal component analysis), как показано на рис. 1. Из рис. 1 видно, что отображение с помощью МГК данных в трехмерном пространстве позволяет предположить возможность группировки данных в кластеры, даже в тех случаях, когда кластеры не имеют строгой геометрической формы.

Существуют несколько зарекомендовавших себя методов («локтя» [7] и «силуэтов» [8]), позволяющих определить значения входных параметров алгоритмов. С их помощью возможно косвенно оценить значения входных параметров алгоритмов DBSCAN (ε и minPts) и OPTICS (minPts) – через определение количества получаемых кластеров k. Па-

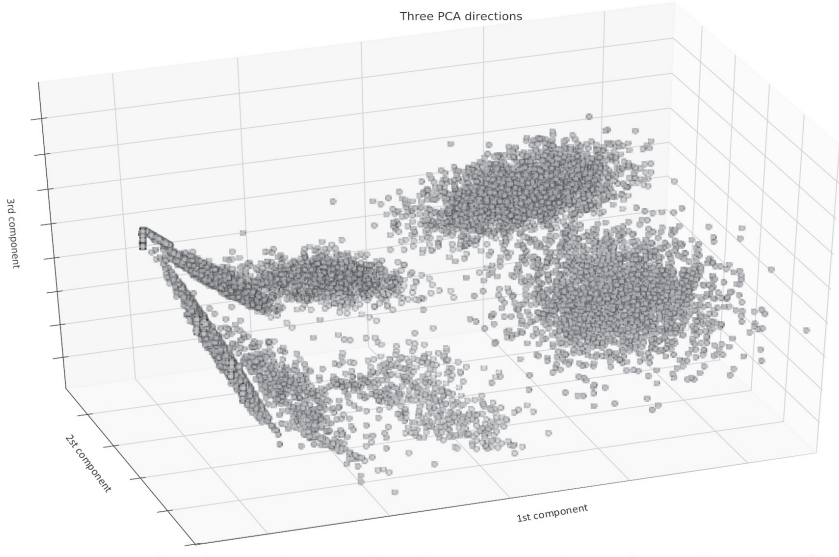


Рис. 1. Пример трехмерного сечения пространства $\mathcal{B}^{1460}$, полученного с использованием МГК

раметр ε – максимальный «размер» окрестности точки, в которой осуществляется поиск ее «соседей». minPts – минимальное число точек в ε окрестности, необходимых для продолжения формирования текущего кластера. Если точки корректно распределены в нужное количество кластеров, то полученные значения входных параметров удовлетворяют условию решаемой задачи группировки. В то же время, данные методы не позволяют автоматизировать процесс подбора входных параметров алгоритма без участия специалиста.

Учитывая указанный недостаток существующих методов определения оптимального значения k, ниже предлагается рассмотреть «принцип минимальной длины описания» (minimum description length – MDL) [9]. Идея MDL заключается в следующем: «Любая закономерность в заданном наборе данных может быть использована для сжатия данных, то есть описания данных с использованием меньшего набора символов, чем нужно для описания данных буквально» [10]. В роли количественной оценки величины, которая необходима для описания набора данных, выступает длина описания L(x). В соответствии с MDL, нахождение минимального значения L(x) позволяет определить оптимальный способ описания данных. Под набором данных понимаются точки на плоскости, связанные с пакетами CT, а закономерностью является принадлежность точек кластерам. В соответствии с MDL, ожидаемым результатом является представление точек на плоскости кластерами, количество которых меньше, чем точек.

В работах [10, 11] показано, что расчет L(x) исследуемых случайно распределённых величин с помощью используемой модели (в рамках исследования моделью является разбиение точек на кластеры) осуществляется по формуле (1):

$$L(x) = -\sum_{x \in X} \log_2 p(x) + \frac{1}{2} P \log_2 |X| \quad (1)$$

Первым слагаемым является отрицательное значение логарифмической функции правдоподобия [12, 11] для всех x ∈ X, где X – множество значений, преобразованных к одномерным скалярным величинам, которые являются компонентами используемой модели. Второе слагаемое – «штраф» за увеличение количества оцениваемых параметров P модели.

Для того, чтобы рассчитать оптимальное значения k необходимо преобразовать дискретно распределенные в пространстве точки к одномерным скалярным величинам. Авторы, по аналогии с работой [13], предлагают воспользоваться расчетом нормы вектора между точкой в пространстве и геометрическим центром кластера. Далее расчет производится с учетом изменений в формуле (1) [14]:

$$L(T', C) = -\sum_{n=1}^{N} \log_2 p(\|T'_n - C\|) + k \log_2 |\mathbb{T}|, \quad (2)$$

где T'_n – точка, ассоциированная с вектором T_n; C – центральная точка кластера (определенная с помощью алгоритма k-медоидов[1]), которому T'_n принадлежит; $p(\|T'_n - C\|)$ – плотность распределения вероятности

[1] Применение алгоритма k-медоидов [15] к сформированному с помощью DBSCAN/OPTICS кластеру позволит определить его центральную точку.

(probability density function – PDF) нормы вектора между T'_n и C; $|\mathbb{T}|$ – количество анализируемых точек.

В ходе экспериментов расчет L(T', C) проводился анализ по оценке применимости различных функций плотности распределения вероятностей и были выбраны оптимальные для применения в MDL быстроубывающие при удалении от центра рассеивания функции плотности распределения вероятности, в отличие от работы [13], где рассматривалось только Гамма-распределение. Дополнительно проведен анализ оптимальных границ диапазонов значений параметров распределений (в рамках решаемой авторами задачи), как указано в табл. 1.

Таблица 1

Значения параметров плотности распределения, используемых при определении оптимального количества кластеров k

Тип распределения	Параметр	Диапазон значений параметра / Оптимальное значение параметра
Нормальное распределение	μ	0-2 / 0
	σ	0.5-1.5 / 1
Гамма-распределение	α	0.5; 1; 2 / 1
	β	0.8-1.1 / 0.9
Распределение χ^2	ν	1-5 / 2

Согласно MDL, оптимальное значение k соответствует минимальному значению L(T', C), рассчитанному по формуле (2). По результатам нахождения оптимального значения k к точкам T'_n применяются выбранные ранее алгоритмы кластеризации DBSCAN или OPTICS с теми значениями входных параметров ε и minPts, при которых получено минимальное . Это возможно осуществить путем перебора параметров алгоритмов DBSCAN (ε и minPts) и OPTICS (только minPts), и на каждом этапе для полученного набора кластеров производить вычисление L(T', C).

Для осуществления перебора параметров алгоритмов необходимо ввести для них граничные значения. Нижней и верхней границами параметра ε будут являться соответственно минимальное и максимальное расстояние между точками T'_n. minPts = 1 не имеет смысла, иначе любая точка будет кластером, следовательно minPts ∈ [2, $|\mathbb{T}|$].

Прерывание цикла перебора значений

Алгоритм № 1: анализ

Input: PDF parameters $\big((\mu,\sigma)|(\alpha,\beta)\big)$, Data ($\mathbb{T}$), Points Amount ($N$)
Result: Clusterized Data (C^*)
1. $\varepsilon_{start}, \varepsilon_{stop}, \varepsilon_{step}$ = processDistances($\mathbb{T}$)
2. $minPts_{start} = 2, minPts_{stop} = N$
3. C^* = clustrize$\Big(\mathbb{T}, minPts_{start}, minPts_{stop}, \varepsilon_{start}, \varepsilon_{stop}, \varepsilon_{step}, (\mu,\sigma)|(\alpha,\beta)\Big)$
4. **return** C^*

Алгоритм № 2: функция «processDistances»

Input: Data ($\mathbb{T}$)
Result: $\varepsilon_{start}, \varepsilon_{stop}, \varepsilon_{step}$
1. $\varepsilon_{start} = \|T'_1 - T'_2\|, \varepsilon_{stop} = 0.0, \varepsilon_{step} = 1.0$
2. **foreach** $T'_n \in \mathbb{T}$ **do**
3. **foreach** $T'_j \in \mathbb{T} \wedge T'_j \neq T'_n$ **do**
4. $d = \|T'_n - T'_j\|$
5. **if** $d > \varepsilon_{stop}$ **then**
6. $\varepsilon_{stop} = d$
7. **end**
8. **if** $d < \varepsilon_{start}$ **then**
9. $\varepsilon_{start} = d$

Окончание таблицы (Алгоритм №2)

```
            end
        end
    end
    if ε_stop < 2.0 then
        ε_step = 0.1
    end
    else if ε_start > 2.0 then
        ε_step = 1.0
    end
    return ε_start, ε_stop, ε_step
```

Алгоритм № 3: функция «clusterize» (for DBSCAN)

Input: Data ($\mathbb{T}$), $minPts_{start}$, $minPts_{stop}$, ε_{start}, ε_{stop}, ε_{step}, PDF Parameters$\big((\mu,\sigma)|(\alpha,\beta)\big)$
Result: Clusterized Data (C^*)

```
foreach ε = overline(ε_start, ε_stop, ε_step) do
    foreach minPts = overline(minPts_start, minPts_stop) do
        O = DBSCAN(𝕋, ε, minPts)
        if ∀ o ∈ O o = noise then
            break
        end
        C = centroids (O, 𝕋, minPts_stop)
        L(T', C) = countDL(O, C, 𝕋, (μ, σ)|(α, β))
        𝕃 ← {L(T', C), ε, minPts}
    end
end
MDL = min(𝕃)
ε, minPts ← MDL
C* = DBSCAN(𝕋, ε, minPts)
return C*
```

Алгоритм № 4: функция «centroids»

Input: Points Cluster Labels (O), Data ($\mathbb{T}$), Points Amount ($minPts_{stop}$)
Result: Cluster Centroids Array (C)

```
foreach cNum ∈ unique(O) do
    foreach pointIndex ∈ overline(0, minPts_stop) do
        if O(T'_pointIndex ∈ 𝕋) = cNum then
            mergedPointsArray← T'_pointIndex
        end
    end
    C ←(cNum, KMedoids(nClusters = 1, mergedPointsArray))
end
return C
```

minPts осуществляется в том случае, если алгоритм кластеризации определил все точки как шум. Для процедуры определения значений параметров DBSCAN выход из цикла перебора minPts приводит к увеличению ε, а для OPTICS – к окончанию процедуры.

Тестирование метода проводилось с использованием типовых наборов данных с известным количеством кластеров (рис. 4). Результаты работы MDL (реальное количество кластеров – количество кластеров с применением оценки MDL) представлены в табл. 2.

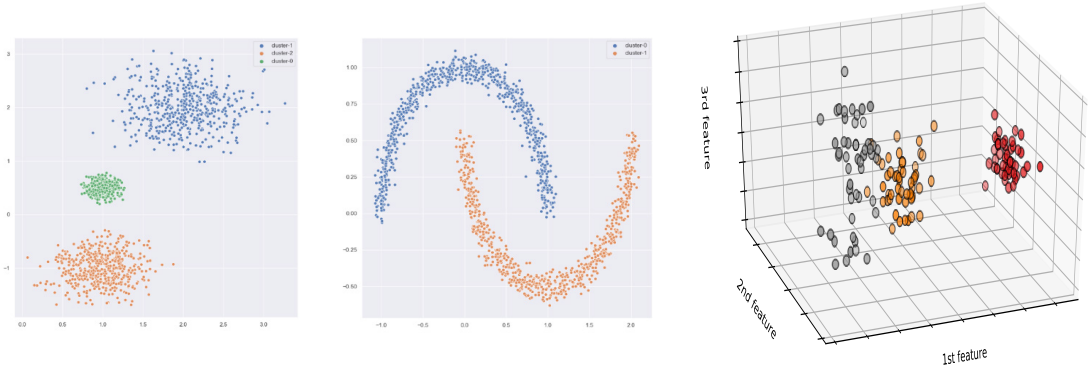
Предложенный метод не во всех случаях позволяет достичь абсолютной точности определения оптимальных параметров кластеризации, особенно когда точки «разреже-

Алгоритм № 5: функция «countDL»

Input: Points Cluster Labels (O), Cluster Centroids Array (C), Data ($\mathbb{T}$), Points Amount ($minPts_{stop}$), PDF Parameters$\big((\mu,\sigma)|(\alpha,\beta)\big)$

Result: Description Length $L(T',C)$

1. $likelyHoodFunctionValue = 0$
2. **foreach** $pointIndex \in \overline{0, minPts_{stop}}$ **do**
3. $cNum = O(T'_{pointIndex} \in \mathbb{T}^*)$
4. $value = \log\left(\text{PDF}\left(\left\|T'_{pointIndex} - C(cNum)\right\|, (\mu,\sigma)|(\alpha,\beta)\right)\right)$
5. $likelyHoodFunctionValue = likelyHoodFunctionValue + value$
6. **end**
7. $L(T',C) = -likelyHoodFunctionValue + \text{size}(C)*\|\mathbb{T}\|$
8. **return** $L(T',C)$



а) б) в)

Рис. 4. Типовые наборы данных: а) Blobs, б) NoisyMoons, в) Iris

Таблица 2

Определение количества кластеров с применением MDL

Набор данных	Кол-во кластеров	DBSCAN, точек		OPTICS, точек	
		150	1500	150	1500
Blobs	3	3	3	3	3
NoisyMoons	2	**5**	2	**4**	2
Iris	2	2		2	

ны» в пространстве. Примеры некорректных значений выделены жирным шрифтом в табл. 2.

3. Применение принципа минимальной длины описания для метода отбора пакетов СТ

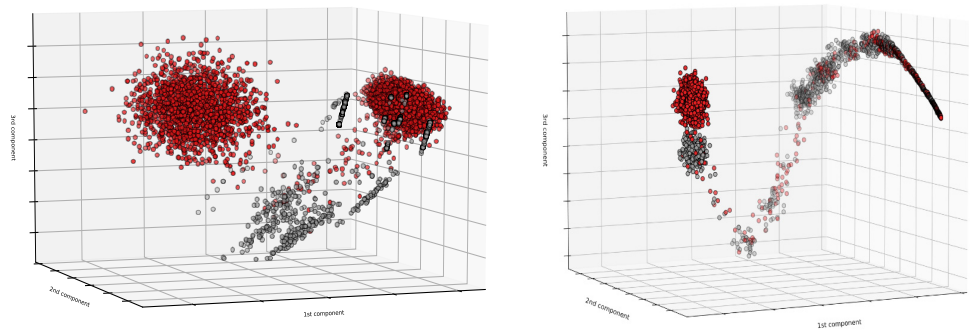
Предложенная оценка входных параметров алгоритмов кластеризации через расчет оптимального k с использованием MDL применена для метода отбора пакетов СТ с целью их группировки и последующей маркировки.

Входными данными являются точки T'_n, соответствующие векторам $T_{n'}$, описывающим пакеты СТ, содержащего в том числе пакеты мессенджеров (пример трехмерных сечений пространства $\mathcal{B}^{1460}$ для мессенджеров WhatsApp и Discord на рис. 5).

Результаты работы метода для алгоритма DBSCAN приведены в табл. 3, а для OPTICS – в табл. 4.

Из значений табл. 4 и 5 видно, что MDL для алгоритмов OPTICS и DBSCAN равен 25,09, при этом точки T'_n были объединены в один кластер. Разбиение точек на два кластера, совпадающее с количеством классов в используемом наборе данных, достигается при L(x) равном 37,64.

Полученная группировка точек на кластеры в последующем должна быть промаркирована специалистом, т.е. соотнесена с классами – мессенджерами. Так, исходя из данных, представленных на рис. 5 и в табл. 4, видно, что сгруппированные в 2 кластера точки (при значении L(x) равном 37,64) мар-



а) б)

Рис. 5. Примеры отображения трехмерных сечений пространства $\mathcal{B}^{1460}$ (точки T'_n, содержащие пакеты мессенджеров WhatsApp и Discord)

Таблица 3

Результаты MDL для оценки работы DBSCAN

L(x)	ε	minPts	Кол-во кластеров
25,09	4	303	1 и шум[2]
37,64	3	298	2 и шум
50,19	3	301	3 и шум
62,73	3	304	4 и шум
75,28	3	306	5 и шум
87,82	3	307	6 и шум
100,37	3	311	7 и шум
112,92	3	314	8 и шум
…			

Таблица 4

Результаты MDL для оценки работы OPTICS

L(x)	minPts	Кол-во кластеров
25,09	5 – 161	1 и шум
25,09	351-650	1 и шум
37,64	161 – 350	2 и шум
100,37	4	7 и шум
263,47	3	20 и шум
2802,63	2	219 и шум

кируются следующим образом: левая и правая верхние группы («шарообразные» кластеры) маркируются как класс WhatsApp, а нижняя группа (кластер произвольной формы) – как Discord.

4. Ускорение подсчета MDL за счет уменьшения признакового пространства

Большое количество признаков снижает скорость кластеризации, анализа данных, а также может отрицательно повлиять на точность результатов [3]. Возникает необходимость обоснованного выбора признаков в целях снижения размерности пространства $\mathcal{B}^{1460}$. Применение алгоритма XGBoost [16] при решении задачи классификации позволило построить диаграмму значимости признаков (рис. 6). Расчет метрики F-мера [17] для построенных моделей с разным количеством признаков от 1460 до 3, отсортированных по важности, приведен в таблице 5. Из расчетов видно, что значимыми для обучения модели являются h*=17 признаков из 1460 (для мессенджера WhatsApp показатели метрики повысились, а для остальных уменьшение размерности признакового пространства не повлияло на F-мера).

Таким образом, вектор T преобразуется в T*, содержащий 17 признаков (список вы-

[2] Разрозненные точки в пространстве, не отнесенные ни к одному кластеру

бранных признаков: 4, 5, 6, 8, 10, 11, 295, 452, 456, 863, 1205, 1399, 1404, 1407, 1412, 1435, 1439). Полученное множество векторов $\mathbb{T}^* = \{T_n^* | n = \overline{1, N}\}$ будет являться входными данными для последующего отбора пакетов СТ.

Результаты работы метода в пространстве $\mathcal{B}^{17}$ для алгоритма DBSCAN приведены в табл. 6, а для OPTICS – в табл. 7.

Из значений табл. 3 и 6 видно, что MDL для алгоритма DBSCAN в пространствах $\mathcal{B}^{17}$ и $\mathcal{B}^{1460}$ равен 37,64, полученное количество кластеров соответствует количеству мессенджеров в трафике, при этом скорость подсчета MDL для пространства, содержащего 17 признаков, на порядок выше.

Применение MDL к алгоритму OPTICS в пространстве $\mathcal{B}^{17}$ (табл. 7) показывает худшие

Рис. 6. График распределения признаков по степени их значимости

Таблица 5

Результаты расчета метрики F-мера для моделей с разным количеством признаков

Кол-во признаков	Мессенджер			
	Discord	WhatsApp	Skype	Шум (https)
3	0,97	0,87	0,89	0,78
…				
7	0,97	0,89	0,92	0,84
8	0,97	0,9	0,91	0,82
9	0,97	0,9	0,92	0,83
10	0,97	0,9	0,92	0,82
11	0,97	0,91	0,91	0,82
…				
16	0,97	0,91	0,92	0,83
17	**0,97**	**0,92**	**0,92**	**0,83**
18	0,97	0,91	0,92	0,83
19	0,97	0,91	0,92	0,83
…				
30	0,97	0,91	0,91	0,82
…				
1460	0,97	0,91	0,92	0,85

Таблица 6

Результаты MDL для оценки работы DBSCAN в признаковом пространстве $\mathcal{B}^{17}$

L(x)	ε	minPts	Кол-во кластеров
37,64	111	5	2 и шум
50,19	106	4	4
50,19	111	4	4
62,73	113	4	5
547,59	1	118	3
554,3	1	117	3
576,27	116	6	3

Таблица 7

Результаты MDL для оценки работы OPTICS в признаковом пространстве $\mathcal{B}^{17}$

L(x)	minPts	Кол-во кластеров
109518,64	32	24
112777,21	3	720
113383,46	26	32
115623,88	27	29
119265,05	4	486
124971,38	33	25
129084,96	31	26

результаты в сравнении с DBSCAN. Это объясняется тем, что процедура определения ε в OPTICS завершается раньше, чем достигается минимальное значение расстояния между точками T'_n в наборе данных.

Исходя из представленного количества точек, маркировка кластеров существенно сокращает временные затраты на формирование «обучающей выборки» в сравнении с маркировкой каждого пакета по отдельности.

Заключение

Предложенный метод группировки пакетов СТ (соотнесенных через векторы T_n, взаимно-однозначно отображенные в точки T'_n) в разных признаковых пространствах позволяет автоматизировать процесс их отбора в целях маркировки для формирования «обучающей выборки». В последующем обученный на основании такой выборки классификатор позволит проанализировать данные в тех признаковых пространствах, которые не могут быть представлены для «ручного» анализа специалистом, но наблюдаются при определении несанкционированного использования мессенджеров.

Литература

1. Большинство российских компаний используют мессенджеры для рабочих вопросов [Электронный ресурс]. URL: https://new-retail.ru/novosti/retail/bolshinstvo_rossiyskikh_kompaniy_ispolzuyut_messendzhery_dlya_rabochikh_voprosov3492. (дата обращения: 01.08.2022).

2. Shabtai A., Elovici Y., Rokach L. A Survey of Data Leakage Detection and Prevention Solutions. Boston, Springer, 2021.

3. Kollerov A.S., Pyr'ev M.S., Fartushnyy A.V. Analysis of Node Interaction Using the TLS Protocol by Means of Machine Learning Tools // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT 2020), P. 524–526.

4. RFC 1191 - Path MTU Discovery [Электронный ресурс]. URL: https://datatracker.ietf.org/doc/html/rfc1191. (дата обращения: 01.08.2022).

5. Ester M., Kriegel H.P., Sander J., Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, P. 226–231.

6. Ankerst M., Breunig M.M., Kriegel H.P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure // ACM SIGMOD International Conference on Management of Data (1999), P. 49–60.

7. Thorndike R.L. Who belongs in the family? // Psychometrika, vol. 18 (1953), P. 267–276.

8. Rousseeuw P.J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis // Computational and Applied Mathematics, vol. 20 (1987), P. 53–65.

9. Rissanen J. Modeling by shortest data description // Automatica, vol. 14 (1978), P. 465–471.

10. Rissanen J. A Universal Prior for Integers and Estimation by Minimum Description Length // The Annals of Statistics, vol. 11 (1983), no. 2, P. 416–431.

11. Roberts S. J. Novelty Detection using Extreme Value Statistics // IEE Proceedings – Vision Image and Signal Processing, vol. 146 (1999), no. 3, P. 124–129.

12. Fisher R. On the mathematical foundations of theoretical statistics // Philosophical Transactions of the Royal Society, vol. 222 (1922), P. 309–368.

13. Using minimum description length to optimize the 'k' in k-medoids. [Электронный ресурс]. URL: http://erikerlandson.github.io/blog/2016/08/03/x-medoids-using-minimum-description-length-to-identify-the-k-in-k-medoids. (дата обращения: 21.03.2022).

14. Гибилинда Р.В. Кластеризационный метод идентификации воздействий на файлы с применением алгоритма k-средних, используемый при расследовании инцидентов информационной безопасности // Вестник УрФО. Безопасность в информационной сфере. – 2020. – Вып. 35 – № 1. – С. 35–47.

15. Kaufman L., Rousseeuw P.J. Clustering by means of medoids // Statistical Data Analysis Based on the L1-Norm and Related Methods, North-Holland, Elsevier (1987), P. 405–416.

16. XGBoost [Электронный ресурс]. https://github.com/dmlc/xgboost. (дата обращения: 01.08.2022).

17. Rijsbergen C.J. Information Retrieval. 2nd Edition. Butterworth, 1979.

References

1. Bol'shinstvo rossiyskikh kompaniy ispol'zuyut messendzhery dlya rabochikh voprosov [Elektronnyy resurs]. URL: https://new-retail.ru/novosti/retail/bolshinstvo_rossiyskikh_kompaniy_ispolzuyut_messendzhery_dlya_rabochikh_voprosov3492. (data obrashcheniya: 01.08.2022).

2. Shabtai A., Elovici Y., Rokach L. A Survey of Data Leakage Detection and Prevention Solutions. Boston, Springer, 2021.

3. Kollerov A.S., Pyr'ev M.S., Fartushnyy A.V. Analysis of Node Interaction Using the TLS Protocol by Means of Machine Learning Tools // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT 2020), P. 524–526.

4. RFC 1191 - Path MTU Discovery [Elektronnyy resurs]. URL: https://datatracker.ietf.org/doc/html/rfc1191. (data obrashcheniya: 01.08.2022).

5. Ester M., Kriegel H.P., Sander J., Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, P. 226–231.

6. Ankerst M., Breunig M.M., Kriegel H.P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure // ACM SIGMOD International Conference on Management of Data (1999), P. 49–60.

7. Thorndike R.L. Who belongs in the family? // Psychometrika, vol. 18 (1953), P. 267–276.

8. Rousseeuw P.J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis // Computational and Applied Mathematics, vol. 20 (1987), P. 53–65.

9. Rissanen J. Modeling by shortest data description // Automatica, vol. 14 (1978), P. 465–471.

10. Rissanen J. A Universal Prior for Integers and Estimation by Minimum Description Length // The Annals of Statistics, vol. 11 (1983), no. 2, P. 416–431.

11. Roberts S. J. Novelty Detection using Extreme Value Statistics // IEE Proceedings – Vision Image and Signal Processing, vol. 146 (1999), no. 3, P. 124–129.

12. Fisher R. On the mathematical foundations of theoretical statistics // Philosophical Transactions of the Royal Society, vol. 222 (1922), P. 309–368.

13. Using minimum description length to optimize the 'k' in k-medoids. [Elektronnyy resurs]. URL: http://erikerlandson.github.io/blog/2016/08/03/x-medoids-using-minimum-description-length-to-identify-the-k-in-k-medoids. (data obrashcheniya: 21.03.2022).

14. Gibilinda R.V. Klasterizatsionnyy metod identifikatsii vozdeystviy na fayly s primeneniem algoritma k-srednikh, ispol'zuemyy pri rassledovanii intsidentov informatsionnoy bezopasnosti // Vestnik UrFO. Bezopasnost' v informatsionnoy sfere. – 2020. – Vyp. 35 – № 1. – C. 35-47.

15. Kaufman L., Rousseeuw P.J. Clustering by means of medoids // Statistical Data Analysis Based on the L1-Norm and Related Methods, North-Holland, Elsevier (1987), P. 405–416.

16. XGBoost [Elektronnyy resurs]. https://github.com/dmlc/xgboost. (data obrashcheniya: 01.08.2022).

17. Rijsbergen C.J. Information Retrieval. 2nd Edition. Butterworth, 1979.

ГИБИЛИНДА Роман Владимирович, кандидат технических наук, доцент учебно-научного центра «Информационная безопасность», Институт радиоэлектроники и информационных технологий-РтФ Уральский федеральный университет им. первого Президента России Б.Н. Ельцина. Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: r.v.gibilinda@urfu.ru

КОЛЛЕРОВ Андрей Сергеевич, кандидат технических наук, доцент, доцент учебно-научного центра «Информационная безопасность» Институт радиоэлектроники и информационных технологий-РтФ Уральский федеральный университет им. первого Президента России Б.Н. Ельцина. Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: a.s.kollerov@ urfu.ru

ФАРТУШНЫЙ Андрей Владимирович, ассистент учебно-научного центра «Информационная безопасность» Институт радиоэлектроники и информационных технологий-РтФ Уральский федеральный университет им. первого Президента России Б.Н. Ельцина. Россия, 620002, г. Екатеринбург, ул. Мира, 19. E-mail: a.v.fartushnyi@urfu.ru

GIBILINDA Roman Vladimirovich, candidate of technical sciences, Associate Professor of Educational and Scientific Center "Information Security", Institute of Radio electronics and Information Technologies, Ural Federal University named after first President of Russia B.N. Yeltsin. 620002, Yekaterinburg, Mira str., 19. E-mail: r.v.gibilinda@urfu.ru.

KOLLEROV Andrey Sergeevich, candidate of technical sciences, Associate Professor, Associate Professor of Educational and Scientific Center "Information Security", Institute of Radio electronics and Information Technologies, Ural Federal University named after first President of Russia B.N. Yeltsin. 620002, Yekaterinburg, Mira str., 19. E-mail: a.s.kollerov@ urfu.ru.

FARTUSHNYI Andrey Vladimirovich, assistant of Educational and Scientific Center "Information Security", Institute of Radio electronics and Information Technologies, Ural Federal University named after first President of Russia B.N. Yeltsin. 620002, Yekaterinburg, Mira str., 19. E-mail: a.v.fartushnyi@ urfu.ru.