Научная статья

УДК 004.056
DOI 10.25205/1818-7900-2022-20-4-61-75

Метод стеганоанализа JPEG-изображений на основе цепей Маркова и его применение в сочетании с различными алгоритмами машинного обучения

Александра Владимировна Прокофьева[1]
Алексей Николаевич Шниперов[2]

[1,2] Сибирский федеральный университет
Красноярск, Россия

[1]prokofe-aleksandra@yandex.ru, https://orcid.org/0000-0001-5104-4511
[2]ashniperov@sfu-kras.ru, https://orcid.org/0000-0003-4231-9805

Аннотация

В работе предложен метод нахождения вектора характеристик изображений, позволяющий эффективно детектировать наличие скрытой информации в изображениях формата JPEG, встроенной различными популярными инструментами стеганографии. Данный метод основан на использовании матриц переходных вероятностей. Кроме того, в работе выполнена сравнительная оценка применения различных технологий машинного обучения для решения задачи стеганоанализа статических изображений формата JPEG, а именно: деревьев решений с градиентным бустингом, линейных моделей, метода k-ближайших соседей, метода опорных векторов, нейронных сетей и искусственных иммунных систем. Приведены результаты качества классификации каждым из вышеперечисленных методов. Сущность метода нахождения вектора характеристик изображения заключается в использовании матрицы переходных вероятностей и применении метода калибровки изображения для повышения точности стеганоанализа и уменьшения числа ложных срабатываний. Для каждого изображения из обучающей и тестовой выборки таким способом находится вектор его характеристик, число элементов которого составляет 324. Далее на полученных данных из обучающей выборки производилось обучение моделей каждым из вышеперечисленных методов машинного обучения отдельно. Тестирование качества построенной модели осуществлялось на данных тестовой выборки также для каждого алгоритма отдельно. Для оценки качества моделей использовались следующие метрики: точность, величина ошибки первого и второго рода результатов бинарной классификации, а также время классификации одного изображения. Для обучения и тестирования методов была использована выборка изображений IStego100K, состоящая из 208 тысяч изображений одинакового размера 1024 × 1024 с различными значениями качества JPEG из диапазона от 75 до 95. Для встраивания скрытого сообщения использовался один из трех алгоритмов стеганографии: J-UNIWARD, nsF5 и UERD. Результатом проведенного исследования является подтверждение того, что предложенный подход нахождения вектора характеристик изображения позволяет детектировать наличие скрытого вложения в изображениях, полученных в результате применения неадаптивных методов стеганографии (Steghide, OutGuess и nsF5) с очень высокой точностью, более 95 %. Для заполненных контейнеров, полученных в результате встраивания сообщения одним из адаптивных методов (J-UNIWARD, UERD), показатели точности обнаружения находятся в пределах 50–60 %. Практическая значимость заключается в экспериментальных данных, подтверждающих эффективность метода стеганоанализа в отношении детектирования скрытой информации в изображениях формата JPEG. Результаты работы могут быть полезны исследователям в области стеганографии и стеганоанализа для сравни-

© Прокофьева А. В., Шниперов А. Н., 2022

тельного анализа применения технологий машинного обучения для решения задачи обнаружения наличия скрытого вложения в изображениях формата JPEG.

Ключевые слова

стеганоанализ, стеганография J-UNIWARD, nsF5, UERD, бинарная классификация, метод k-ближайших соседей, метод опорных векторов, нейронные сети, искусственные иммунные системы

Для цитирования

Прокофьева А. В., Шниперов А. Н. Метод стеганоанализа JPEG-изображений на основе цепей Маркова и его применение в сочетании с различными алгоритмами машинного обучения // Вестник НГУ. Серия: Информационные технологии. 2022. Т. 20, № 4. С. 61–75. DOI 10.25205/1818-7900-2022-20-4-61-75

A Markov Chain – Based Method for JPEG Image Steganalysis and Its Application in Combination with Various Machine Learning Algorithms

Aleksandra V. Prokofieva[1]
Alexey N. Shniperov[2]

[1,2]Siberian Federal University
Krasnoyarsk, Russian Federation

[1]prokofe-aleksandra@yandex.ru, https://orcid.org/0000-0001-5104-4511
[2]ashniperov@sfu-kras.ru, https://orcid.org/0000-0003-4231-9805

Abstract

The paper proposes a method of extracting the feature vector of images, which makes it possible to effectively detect the presence of hidden information in JPEG images embedded by various popular steganography tools. This method is based on the usage of the transition probability matrix. The essence of the method for extracting the feature vector of the image is to use the transition probability matrix and apply the image calibration method to improve the accuracy of steganalysis and reduce the number of false positives. For each image from the training and test sets a feature vector is found in this way, the number of elements is 324. Further, the models were trained on the training dataset by each of machine learning methods separately: decision trees with gradient boosting, linear models, k-nearest neighbors, support vector machines, neural networks, and artificial immune systems. To assess the capacity of the models the following metrics were used: accuracy, the rate of the false positive and false negative errors, and the confusion matrix. The results of classification by each of the above methods are given. For training and testing a dataset IStego100K was used, which consists of 208 thousand images of the same size 1024 × 1024 with different quality values in the range from 75 to 95. One of the J-UNIWARD, nsF5, and UERD steganography algorithms was used to embed a hidden message. As a result, we can observe that the proposed approach to extracting the feature vector makes it possible to detect the presence of hidden information embedded by non-adaptive steganography (Steghide, OutGuess and nsF5) in static JPEG images with high accuracy (more than 95%). However, for adaptive steganography methods (J-UNIWARD, UERD) the accuracy is less (about 50-60%).

Keywords

steganoanalysis, steganography, J-UNIWARD, nsF5, UERD, binary classification, k-nearest neighbors, support vector machines, neural networks, artificial immune systems

For citation

Prokofieva A. V., Shniperov A. N. A Markov Chain – Based Method for JPEG Image Steganalysis and Its Application in Combination with Various Machine Learning Algorithms. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 4, pp. 61–75. (in Russ.) DOI 10.25205/1818-7900-2022-20-4-61-75

Введение

Большая часть последних исследований в области стеганографии и стеганоанализа сосредоточилась на изображениях формата JPEG ввиду широкого использования этого формата

для передачи изображений в сети Интернет. Высокая популярность данного формата изображений привела к появлению множества новых методов скрытой передачи информации, в которых в качестве стеганоконтейнера выступает изображение. Такие стеганоконтейнеры могут содержать в себе противоправную информацию (к примеру, экстремистского или криминального характера), также с помощью таких методов возможно обойти мониторинг средств защиты информации, таких как средств антивирусной защиты и средств предотвращения утечек конфиденциальной информации (DLP-систем). Ввиду всего вышесказанного можно говорить о высокой актуальности задачи стеганоанализа изображений формата JPEG, развитии новых методов обнаружения скрытых каналов передачи информации.

На сегодняшний день существует большое количество методов стеганоанализа, которые различаются по анализируемым характеристикам изображения и методам встраивания, которым они противодействуют. Методы стеганоанализа традиционно разделяют на сигнатурные, статистические и эвристические. Также можно отдельно выделить четвертый тип: метод визуального поиска, который предполагает визуальный просмотр каждого изображения, а также различные преобразования изображения в пространственной области (преобразование яркости, цветовая коррекция и т. д.). Однако поскольку этот подход сильно зависим от того изображения, в которое производится встраивание скрытого сообщения, и требует непосредственного участия эксперта для анализа изображения, его сложно применять на практике в условиях большого трафика изображений. Сигнатурные методы стеганоанализа предназначены для работы с форматными методами скрытой передачи информации, которые встраивают скрытое сообщение в определенные структурой файла места (например, поле комментария формата файла JPEG) или в процессе встраивания оставляют специфические сигнатуры, по которым удается детектировать скрытое вложение. Статистические методы стеганоанализа основываются на анализе статистических характеристик исследуемого изображения, их корреляции с характеристиками пустых стеганоконтейнеров такого же типа. Наиболее известными статистическими методами являются RS-стеганоанализ и WS-стеганоанализ [1], гистограммный [2] и другие подходы. Данные методы могут показывать очень высокие показатели точности обнаружения заполненных стеганоконтейнеров, а также по определению объема скрытого вложения, однако их точность в значительной степени зависит от метода стеганографии, которым они противодействуют (они могут отлично работать на некоторых методах и совершенно не работать против любых других). Эвристические методы стеганоанализа представляют большой интерес для исследователей, поскольку они не привязаны к какому-то определенному алгоритму встраивания скрытой информации, хоть и в целом не достигают таких показателей, как статистические. Однако для использования на практике, когда алгоритм скрытой передачи информации не известен, они применимы в большей степени. В основном данные методы базируются на решении задачи бинарной классификации с применением методов машинного обучения, например, методы, предложенные в работах [3; 4; 5]. Одной из последних значимых работ является атака обратной совместимости [6], которая основывается на том факте, что изменения в квантованных коэффициентах дискретного косинусного преобразования (ДКП), появившиеся в результате встраивания скрытого сообщения, увеличивают дисперсию гауссова распределения ошибок округления в пространственной области. Этот метод позволяет надежно обнаруживать заполненные стеганоконтейнеры, полученные любым методом стеганографии, даже при небольшом размере скрытого вложения. Однако данный метод работает только на изображениях JPEG с качеством 99 и 100.

В работе [7] нами уже был представлен метод стеганоанализа изображений, основанный на применении искусственных иммунных систем (ИИС), в котором на этапе предобработки изображения вектор характеристик вычислялся с помощью вейвлет-преобразования Хаара, длина такого вектора характеристик для квадратных изображений была равна 36, для прямоугольных – 54. Однако этот метод обладал некоторыми недостатками:

– недостаточная точность классификации изображений (75–80 %). Для применения в реальных информационных системах на практике данный показатель недостаточен, поскольку приводит к высокой вероятности ложных срабатываний и пропусков события передачи изображения, содержащего скрытое вложение;

– продолжительное время обучения искусственной иммунной системы в следующих условиях работы: для обучения и тестирования ИИС использовалась небольшая база изображений, состоящая всего из 7,5 тыс. изображений. Осуществлялось только 10 итераций алгоритма клональной селекции на этапе обучения, в программной реализации не использовались никакие методы распараллеливания вычислений. Поскольку в данной работе для построения метода классификации изображений (после предобработки) использовался только метод на основе искусственных иммунных систем, мы не можем получить полное представление о его эффективности и влиянии на результаты обнаружения изображений – заполненных контейнеров, а также влиянии на время классификации одного изображения.

Поэтому основной целью настоящей работы является исправление существующих недостатков метода стеганоанализа на основе ИИС, а именно: повышение точности классификации изображения путем модификации метода предобработки и получения вектора характеристик изображения, а также проведение сравнительной оценки применения различных иных технологий машинного обучения, благодаря чему можно будет сделать выводы об эффективности применения метода на основе ИИС.

Помимо этого, среди опубликованных статей довольно редко встречаются работы, в которых проводится сравнение методов стеганоанализа на основе машинного обучения. В большинстве случаев оценка эффективности метода и результирующие показатели точности предоставляются непосредственно авторами статей. Проведение сравнительного анализа различных методов стеганоанализа на базе предоставленных авторами показателей довольно затруднительно, поскольку для оценки эффективности методов разными авторами используются различные выборки изображений, методы встраивания скрытого сообщения, различные метрики эффективности алгоритмов, и т. д. Таким образом, проведение сравнения эффективности различных методов стеганоанализа в одинаковых условиях может быть полезно для коллег, занимающихся исследованиями в данной области.

Описание методики проведения экспериментов

Итак, для начала сформулируем общую задачу метода стеганоанализа. Пусть $I = C \cup S$ – множество объектов заданного типа (изображений формата JPEG), S – множество заполненных стеганоконтейнеров, каждый из которых содержит скрытую информацию, C – множество пустых стеганоконтейнеров, не содержащих скрытой информации, полагаем $S \cap C = \varnothing$. Каждый из объектов $img \in I$ представлен вектором $\overline{D}$ его характеристик. Общая постановка задачи стеганоанализа изображения $img \in I$ заключается в решении задачи бинарной классификации наличия/отсутствия скрытого вложения.

Для обучения и тестирования методов классификации нами была использована база изображений IStego100K [8], состоящая из 208 104 изображений одинакового размера 1024 × 1024, все изображения из обучающей и тестовой выборок взяты из одного источника. Среди них 200 тысяч изображений составляют обучающую выборку, а оставшиеся 8 104 – тестовую выборку (значения приведены суммарно для всех алгоритмов и значений коэффициента качества изображения, т. е. для разных алгоритмов используются разные изображения). Для каждого изображения в IStego100K коэффициенты качества JPEG различаются и находятся в диапазоне от 75 до 95. Для встраивания скрытого сообщения использовался один из трех алгоритмов стеганографии: J-UNIWARD, nsF5 и UERD. В базе изображений IStego100K есть различные наборы, отличающиеся величиной стеганографического встраивания: 0,1, 0,2, 0,3, 0,4 bpnz (бит на не-

нулевой AC-коэффициент ДКП), в данной работе для экспериментов использовались заполненные контейнеры, величина стеганографического воздействия в которых составляет 0,4 bpnz.

Алгоритм nsF5 [9] считается неадаптивным к содержимому изображения методом стеганографии, т. е. встраивание скрытого сообщения в изображение происходит последовательно. Но в отличие от других подобных методов, таких как JSteg, встраивание не так сильно влияет на изменение гистограмм коэффициентов ДКП, поэтому данный метод не подвержен основным статистическим атакам, и, тем самым, менее подвержен стеганоанализу.

Алгоритмы J-UNIWARD [10] и UERD [11] являются на данный момент самыми надежными, современными, а также не поддающимися детектированию обычными статистическими методами стеганоанализа. Стеганография с использованием минимальных искажений является наиболее успешной моделью адаптивной стеганографии, при которой сообщение встраивается в наиболее зашумлённые места изображения, и встраивание зависит от сложности текстуры. Алгоритм J-UNIWARD обладает выдающимися характеристиками резистивности к стеганоанализу, в то время как алгоритм UERD имеет меньшую сложность для обнаружения.

Получение вектора характеристик изображения

Этап получения векторов характеристик изображений является одним из ключевых в задаче стеганоанализа и напрямую может определять эффективность того или иного метода.

Поскольку метод, основанный на вычислении вейвлет-преобразования Хаара на практике показывал недостаточную точность классификации, как нам хотелось бы видеть для применения системы в реальных условиях, нами был модифицирован метод, основанный на нахождении матрицы переходных вероятностей изображения (subtractive pixel adjacency matrix), использовавшийся авторами в статьях [5; 9]. Вектор характеристик в таких методах представляет собой марковский процесс как разницу между абсолютными значениями соседних коэффициентов ДКП (дискретного косинусного преобразования), которому, как известно, подвергается каждый блок изображения при сжатии JPEG.

Итак, суть предлагаемого метода заключается в том, что на основе матрицы коэффициентов ДКП изображения *img* размером $N \times M$, согласно формулам (1), находятся четыре разностные матрицы соседних значений коэффициентов ДКП. На рис. 1 представлен принцип получения разностных матриц для горизонтального направления. На этом этапе они будут вычислены для четырех направлений: горизонтального, вертикального и диагонального направлений и направления побочной диагонали, которые обозначаются соответственно $F_h(u, v)$, $F_v(u, v)$, $F_d(u, v)$ и $F_m(u, v)$.

$$F_h(u, v) = F(u, v) - F(u + 1, v), \quad F_v(u, v) = F(u, v) - F(u, v + 1) \tag{1}$$
$$F_d(u, v) = F(u, v) - F(u + 1, v + 1), \quad F_m(u, v) = F(u + 1, v) - F(u, v + 1)$$

где $u \in [1, N - 1]$, $v \in [1, M -]$.

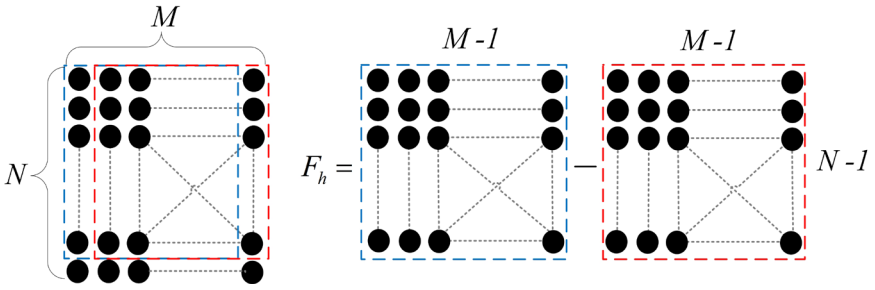

Рис. 1. Принцип получения разностной матрицы $F_h(u, v)$ в горизонтальном направлении
Fig.1. The principle of obtaining a difference matrix $F_h(u, v)$ in the horizontal direction

Далее на основе этих разностных матриц вычисляются матрицы переходных вероятностей, согласно формулам (2). Значение коэффициентов ДКП изображения прогнозируется исходя из значений соседних коэффициентов, а величина ошибки прогнозирования вычисляется путем вычитания значения предсказания из значения исходного коэффициента. Затем происходит сравнение с заданным пороговым значением.

Поэтому, прежде чем найти матрицы переходных вероятностей, сначала нужно задать порог T (примем $T = 4$, поскольку в результате экспериментов это значение позволяет получить лучшие результаты детектирования заполненных стеганоконтейнеров). Значения F_h, F_v, F_d и F_m округляются таким образом, чтобы они попадали в диапазон $[-T, T]$. По полученным значениям F_h, F_v, F_d и F_m рассчитываются матрицы вероятностей перехода $M_h(i,j)$, $M_v(i,j)$, $M_d(i,j)$ и $M_m(i,j)$ следующим образом:

$$M_h(i,j)=\frac{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_h(u,v)-i,F_h(u+1,v)-j\right]}{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_h(u,v)-i\right]},\quad M_v(i,j)=\frac{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_v(u,v)-i,F_v(u,v+1)-j\right]}{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_v(u,v)-i\right]},$$

$$M_d(i,j)=\frac{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_d(u,v)-i,F_d(u+1,v+1)-j\right]}{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_d(u,v)-i\right]}\quad M_m(i,j)=\frac{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_m(u+1,v)-i,F_m(u,v+1)-j\right]}{\sum_{v=1}^{N-1}\sum_{u=1}^{M-1}\delta\left[F_m(u+1,v)-i\right]},\quad (2)$$

где N, M – размеры изображения img, $i, j \in [-T, T]$; δ – дельта (символ) Кронекера (функция двух целых переменных, которая равна 1, если они равны, и 0 в противном случае).

Число элементов вектора характеристик составляет $4 \times (2T + 1)^2$. В нашем случае оно составляет 324, поскольку T выбрано равным 4.

Для повышения эффективности работы метода мы предлагаем использовать метод калибровки – преобразование изображения, позволяющее получить приблизительное отражение статистических свойств пустого стеганоконтейнера для анализируемого изображения. Анализируемое изображение *img* переводится в пространственную область (обратным дискретным косинусным преобразованием). Далее производится его обрезка на четыре пикселя с двух сторон и повторное сжатие полученного изображения с использованием матрицы квантования изображения *img*. В результате получается изображение *img′* формата JPEG, которое будет использоваться в качестве изображения – пустого контейнера. Далее для этого изображения *img′* находится вектор характеристик способом, рассмотренным выше для изображения *img*.

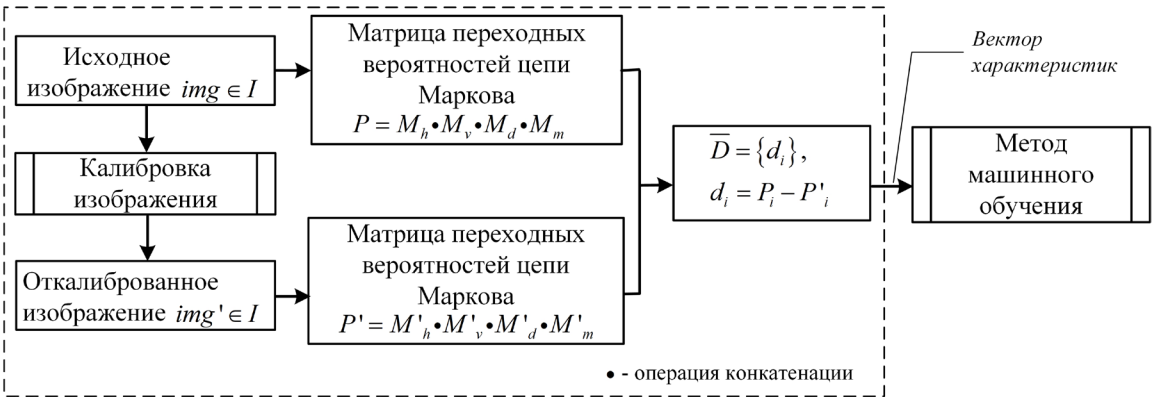

Рис. 2. Принцип получения вектора характеристик изображения
Fig. 2. The principle of obtaining the vector of image characteristics

Таблица 1

Результаты сравнения точности классификации для методов
на основе вейвлет-преобразования Хаара и нахождении матрицы переходных вероятностей

Table 1

Results of Comparison of Classification Accuracy for Methods Based
on the Haar's Wavelet Transform and Finding the Transition Probability Matrix

Алгоритм стеганографии / Метод нахождения вектора характеристик	Steghide	OutGuess	nsF5
Точность обнаружения, %			
Вейвлет-преобразование Хаара [7]	71,1	74,5	74,9
Матрица переходных вероятностей	96,88	97,91	95,57
Величина ошибки I рода, %			
Вейвлет-преобразование Хаара [7]	26,3	16,5	17,1
Матрица переходных вероятностей	3,9	2,3	4,6
Величина ошибки II рода, %			
Вейвлет-преобразование Хаара [7]	31,5	34,5	32,9
Матрица переходных вероятностей	1,92	1,23	3,7

На следующем шаге каждая компонента итогового вектора характеристик изображения находится как разность соответствующих компонент векторов откалиброванного и исходного изображений, поскольку это позволяет снизить зависимость точности бинарной классификации от обучающей выборки изображений.

Итак, на рис. 2 приведена общая схема получения вектора характеристик изображения.

В результате проведенных экспериментов (приведены в табл. 1 и 2) мы можем наблюдать, что данный подход в выделение вектора характеристик изображения позволяет детектировать наличие скрытого вложения в лучшем из неадаптивных методов стеганографии (nsF5) с очень высокой точностью. Результаты сравнения работы данного метода нахождения векторов характеристик изображения и метода, основанного на нахождении вейвлет-преобразования Хаара (предложенного нами в предыдущей работе [7]), приведены в табл. 1. Мы можем видеть, насколько более эффективно данный метод стал работать на неадаптивных алгоритмах стеганографии, таких как nsF5, Steghide, OutGuess: точность обнаружения выросла более чем на 20 %.

Однако на адаптивных методах стеганографии и особенно на J-UNIWARD (который, к слову, хуже всех на данный момент поддается стеганоанализу) результаты пока что оставляют желать лучшего. И над увеличением точности детектирования адаптивных методов стеганографии мы работаем дальше.

Описание рассматриваемых методов машинного обучения

При решении задачи классификации часто возникает сложность с выбором методов машинного обучения, так как их существует достаточно много, а проверить качество работы каждого в отдельности не представляется возможным. При этом каждый из методов обладает своими слабыми и сильными сторонами, имеет свою область применимости.

В данной работе мы провели сравнительную оценку эффективности в задаче стеганоанализа самых популярных алгоритмов машинного обучения: деревьев решений с градиентным

бустингом, линейных моделей (реализованных в библиотеке LightAutoML [13]), метода *k*-ближайших соседей, метода опорных векторов (support vector machines – SVM), нейронных сетей, и искусственных иммунных систем [7].

В общем случае для каждого изображения из обучающей и тестовой выборки находится вектор характеристик (способом, описанным в предыдущем подразделе), состоящий из 324 значений. Далее для каждого алгоритма стеганографии, использованного для получения множества заполненных контейнеров выборки (J-UNIWARD, nsF5 и UERD), отдельно обучается модель (одним из вышеперечисленных методов машинного обучения) на векторах характеристик изображений из обучающей выборки. Завершающим этапом является тестирование качества построенной модели (точность классификации) на данных тестовой выборки.

Опишем подробнее, каким образом настраивались, подбирались параметры каждого из представленных методов в приведенных конкретных условиях и в контексте решения задачи стеганоанализа изображений.

Метод k-ближайших соседей

Как известно, метод *k*-ближайших соседей – алгоритм классификации, где принадлежность объекта к классу определяется за счет наименьшего расстояния между группой объектов известного класса. Ближайшие соседи находятся исходя из множества обучающей выборки, и по ключевому для данного метода значения *k* высчитывается, какой класс наиболее многочислен среди них. За счет вычисления расстояния между всеми объектами выполнение данного алгоритма занимает весьма продолжительное время и требует значительных ресурсов по используемой памяти. Однако данный алгоритм устойчив к выбросам, а единственным неизвестным параметром является *k*.

Для практической реализации данного метода использовалась библиотека scikit-learn на языке Python. Параметр *k* был определен с помощью метода локтя (elbow rule) – для этого мы построили график, отражающий зависимость между значениями *k* и величиной, равной проценту ошибочно классифицированных объектов (сумма величины ошибок первого и второго рода). Точка, где эта величина является минимальной, и считается оптимальным значением *k*. В нашем случае оптимальное значение $k = 2$ (см. рис. 3).

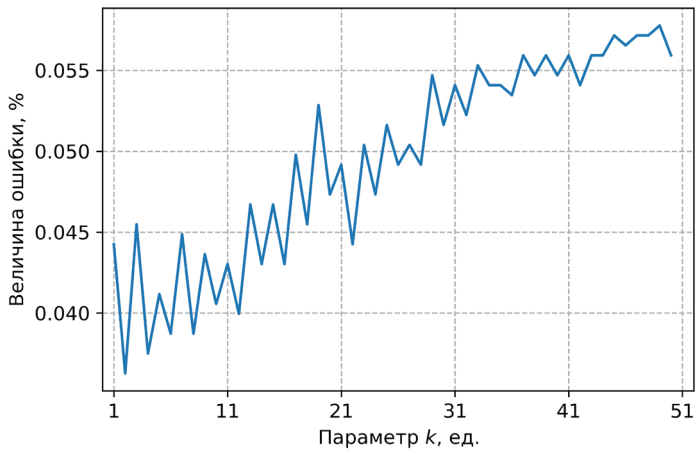

Рис. 3. Зависимость между значением *k* и величиной ошибочно классифицированных объектов тестовой выборки
Fig. 3. The dependence between the value of *k* and the value of the incorrectly classified objects of the test dataset

Деревья решений с градиентным бустингом, линейные модели (библиотека LightAutoML)

LightAutoML – это библиотека на языке Python, написанная сотрудниками лаборатории Artificial Iintelligence Сбербанка, позволяющая автоматически построить модели машинного

обучения [13]. Данная библиотека обеспечивает оптимальный и быстрый поиск гиперпараметров с использованием методов итеративной оптимизации и байесовских методов. У библиотеки LightAutoML в арсенале два метода машинного обучения, которые она моделирует: деревья решений с градиентным бустином (GBMs) и линейные модели. Разработчиками были выбраны именно эти алгоритмы машинного обучения, потому что несмотря на тенденцию в развитии нейронных сетей для разных областей, методы, основанные на GBMs, показывают высокие результаты производительности на табличных данных и превосходят другие подходы во многих тестах и соревнованиях. Кроме того, линейные модели быстры и могут повысить производительность моделей на основе деревьев решений в ансамблях [13].

Для обучения модели потребовалось выполнить несколько настроек. На первом шаге необходимо указать тип модели, в нашем случае это binary – бинарная классификация, поскольку для каждого изображения $img \in I$ требуется определить, относится ли оно к множеству заполненных стеганоконтейнеров S или множеству пустых стеганоконтейнеров C. На втором шаге необходимо импортировать *.csv файлы с векторами характеристик обучающей и тестовой выборок и добавить к ним колонку со значением принадлежности изображения к классу заполненных контейнеров (0 – если является пустым стеганоконтейнером, 1 – если заполненным). И на следующем шаге запустить функцию построения модели, в качестве параметров выбрано использование сочетания обоих вышеперечисленных алгоритмов. После чего остается только протестировать полученную модель на тестовой выборке изображений и оценить результаты (результаты приведены в табл. 2).

Метод опорных векторов (SVM)

Метод (машина) опорных векторов (SVM) – это расширение классификатора опорных векторов, которое является результатом определенного расширения пространства объектов с использованием ядер. Функции «ядра» позволяют вычислять различные разделяющие гиперплоскости, где под «ядром» подразумевается преобразование данных. Данный метод также был реализован с помощью библиотеки sklearn.svm на языке Python. Для экспериментов использовались несколько функций «ядра»: линейное ядро, радиальная базисная функция (RBF), полиномиальное ядро различных степеней, сигмоид-ядро. Наилучшие показатели точности для задачи стеганоанализа позволило получить использование радиальной базисной функции:

$$K(D_i, D'_i) = e^{-\gamma \sum_{n=1}^{L} (D_{ij} - D'_{ij})^2},$$

где D_i, D'_i – векторы характеристик двух изображений, L – длина вектора характеристик (в нашем случае $L = 324$), γ – некоторая положительная константа (для экспериментов использовалась $\gamma = 100$).

Нейронные сети

Нами была сконфигурирована нейронная сеть с помощью библиотеки Keras для Python, которая позволяет максимально ускорить и упростить процесс создания нейронных сетей, путем перебора подбирая наилучшую конфигурацию и параметры нейронной сети.

На первом шаге были импортированы данные обучающей и тестовой выборок и определили последовательный тип модели. Далее протестировали несколько конфигураций. Здесь приведем следующую, которая позволила достичь лучших показателей на тестовой выборке изображений: нейросеть с двумя скрытыми слоями: входной слой состоит из 480 нейронов, и каждый скрытый слой – из 112 нейронов, на выходном слое всего 1 нейрон, поскольку в нашем случае решается задача бинарной классификации.

В качестве функции активации на скрытых слоях использовалась relu (Rectified Linear Unit), на выходном слое использовали сигмоидную функцию, которая выполняет перенормировку значений в диапазоне от 0 до 1.

На этапе компилирования модели использовали следующие параметры: в качестве оптимизатора использовали стохастический градиентный спуск (SGD), так как это позволяет сократить задействованные вычислительные ресурсы и помогает достичь более высокой скорости обучения. В качестве функции потерь используем бинарную кросс-энтропию, в качестве метрики оценки – точность бинарной классификации.

Далее на этапе обучения использовались следующие параметры: фиксированный размер подмножества обучающей выборки (размер пакета, англ. batch size) равный 64 и двадцать эпох обучения (размер пакета определяет количество элементов, которые будут распространяться по сети). Большее количество эпох обучения приводило к переобучению модели и соответственно к худшим результатам на тестовой выборке.

Искусственные иммунные системы

В основе искусственных иммунных систем лежит факт обнаружения и нейтрализация чужеродных объектов (антигенов) иммунной системой. Антигены провоцируют иммунный ответ организма, который начинает вырабатывать защитные клетки иммунной системы – антитела различных видов, пока не найдется такое антитело, которое специфично связывается с антигеном и нейтрализует его. Совокупность сформированных в течение жизни антител формируют иммунитет организма [14]. Применение ИИС для решения задач стеганоанализа является сравнительно новым подходом, однако за последние несколько лет уже опубликован ряд работ в этой области (для стеганоанализа в области JPEG – работа [15]).

Для построения искусственной иммунной системы было принято решение использовать сочетание двух подходов: алгоритма отрицательного отбора, применяющегося при инициализации ИИС, и клонального отбора, применяющегося на этапе обучения ИИС [14], поскольку так мы можем создать систему, которая будет отличать чужеродные клетки от своих, а далее на этапе обучения увеличить относительный размер популяции тех антител, которые доказали свою ценность при распознавании. На данном этапе для обучения использовалось 10 поколений мутации антител. А для получения первоначального набора антител из обучающей выборки выделяются только изображения, являющиеся заполненными контейнерами, дополнительно генерируется небольшое количество антител случайным образом.

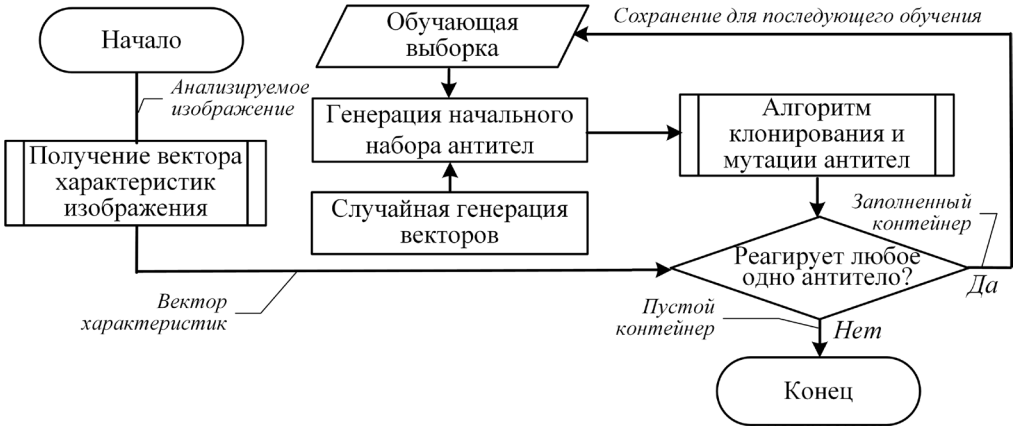

Рис.4. Общая структурно-функциональная схема предлагаемого метода
Fig. 4. General structural and functional scheme of the proposed method

После получения итогового набора антител переходим к фазе тестирования полученной системы, основа работы которой строится на том, что если вектор характеристик анализируемого изображения попадает в окрестность хотя бы одного антитела, то такое изображение бу-

дет отнесено к множеству заполненных стеганоконтейнеров. В противном случае, к множеству пустых.

В общем виде структурно-функциональная схема предлагаемой искусственной иммунной системы для решения задачи стеганоанализа изображений представлена на рис. 4.

Анализ полученных результатов и оценка эффективности методов

Обучение и тестирование ИИС проводились на основе базы изображений IStego100K, состоящей из изображений одинакового размера 1024 × 1024 с различными значениями качества JPEG из диапазона от 75 до 95, как уже говорилось ранее, встраивание скрытого сообщение производилось с помощью алгоритмов стеганографии J-UNIWARD, nsF5 и UERD.

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются различные метрики. В этой статье, поскольку решается задача бинарной классификации, мы будем использовать следующие критерии качества модели: точность (accuracy), величину ошибки первого и второго рода результатов бинарной классификации. Точность показывает количество правильно классифицированных элементов (истинно положительных и истинно отрицательных) от общего числа элементов тестовой выборки. Помимо этого, для оценки качества модели мы будем использовать время, затраченное на классификацию одного изображения на этапе тестирования, поскольку для применения метода на практике в качестве составной части средств защиты информации оно имеет критическое значение (если обработка изображения в высоконагруженной корпоративной сети будет занимать слишком большое время, то, вероятнее всего, от данного функционала системы защиты информации откажутся, ввиду нарушения доступности). Также можно сравнивать время, затраченное на обучение модели, но делать это не совсем корректно, поскольку оно зависит от используемой программной реализации (библиотек или собственноручно написанного кода) и использования параллельных вычислений в данной реализации. Также обучение модели выполняется единожды и не столь критично влияет на выбор алгоритма классификации.

Итак, время классификации одного изображения включает в себя нахождение вектора характеристик цветного изображения размером 1024 × 1024 пикселей описанным выше методом и непосредственно его классификацию одним из алгоритмов классификации.

В табл. 2 представлены результаты метрик точности и величины ошибки первого и второго рода для каждого алгоритма машинного обучения, представленного в данной статье отдельно для каждого алгоритма стеганографии, который использовался для построения выборки изображений.

По результатам проведенных экспериментов можно сделать следующие выводы. Метод *k*-ближайших соседей показывает довольно высокую эффективность, но требует наибольших вычислительных затрат при классификации изображения, следовательно, увеличивает время классификации изображения, что особенно важно при работе в реальных системах с изображениями большого размера.

Метод опорных векторов является очень точным классификатором, устойчив к шуму и наименее предрасположен к переобучению. Нейронные сети применяются в сложных областях и характеризуются большим количеством параметров и показывают очень хорошие результаты. Однако общим недостатком указанных методов является следующий факт: достижение высоких результатов требует серьезных временных затрат на обучение модели.

Метод на основе искусственных иммунных систем также показывает достаточно высокие результаты точности классификации, сравнимые с остальными развитыми методами машинного обучения (такими как нейронные сети, например), однако процесс репродукции антител на этапе обучения также является вычислительно затратным.

Таблица 2

Результаты сравнительной оценки применения предложенного метода нахождения вектора характеристик изображения в сочетании с различными алгоритмами машинного обучения

Table 2

Results of Comparative Evaluation of the Application of the Proposed Method of Finding the Vector of Image Characteristics in Combination with Various Machine Learning Algorithms

	Точность, %	Величина ошибки I рода, %	Величина ошибки II рода, %	Время классификации одного изображения, мс
Алгоритм встраивания скрытого сообщения nsF5				
k-ближайших соседей	95,88	4,9	0,92	240,52
Деревья решений с градиентным бустином (GBMs) + линейные модели (LightAutoML)	97,91	2,3	1,23	210,46
Метод опорных векторов	97,91	2,23	1,54	206,1
Нейронная сеть с 2 скрытыми слоями	97,97	0	3,1	206,39
Искусственные иммунные системы (AIS)	95,57	4,6	3,7	213,5
Алгоритм встраивания скрытого сообщения J-UNIWARD				
k-ближайших соседей	60,68	31,7	69,5	240,52
Деревья решений с градиентным бустином (GBMs) + линейные модели (LightAutoML)	46,2	52,5	59,07	210,46
Метод опорных векторов	52,76	46,73	49,23	206,1
Нейронная сеть с 2 скрытыми слоями	25,6	90,9	8	206,39
Искусственные иммунные системы (AIS)	52,82	45,12	55,6	213,5
Алгоритм встраивания скрытого сообщения UERD				
k-ближайших соседей	64,65	29,7	56,89	240,52
Деревья решений с градиентным бустином (GBMs) + линейные модели (LightAutoML)	33,33	65,6	70,6	210,46
Метод опорных векторов	65,14	36,22	29,62	206,1
Нейронная сеть с 2 скрытыми слоями	34	81,7	4,1	206,39
Искусственные иммунные системы (AIS)	63,2	38,1	31,3	213,5

Заключение

В результате данной работы были улучшены показатели точности классификации изображения путем применения метода нахождения матрицы переходных вероятностей и метода

калибровки изображения, необходимого для оценки вектора характеристик изображения, являющегося пустым стеганоконтейнером. Предложенный подход нахождения вектора характеристик изображения позволяет детектировать наличие скрытого вложения в изображениях, полученных в результате применения неадаптивных методов стеганографии (Steghide, OutGuess и nsF5) с очень высокой точностью более 95 %. Таким образом, точность обнаружения выросла более чем на 20 % по сравнению с методом, использовавшимся нами в предыдущей работе. Для заполненных контейнеров, полученных в результате встраивания сообщения одним из адаптивных методов (J-UNIWARD, UERD), показатели точности обнаружения пока находятся в пределах 50–60 %, что требует доработки в дальнейших исследованиях. Также данный метод нахождения вектора характеристик позволяет классифицировать изображение за очень короткий промежуток времени (до 250 мс), что является вполне приемлемым для использования на практике.

Кроме того, в данной работе была проведена сравнительная оценка применения различных технологий машинного обучения для решения задачи стеганоанализа статических изображений формата JPEG при использовании метода нахождения вектора характеристик на основе использования матрицы переходных вероятностей. Таким образом, метод, основанный на применении искусственных иммунных систем, показывает близкие по точности результаты детектирования в сравнении с классическими методами классификации (методом *k*-ближайших соседей и методом опорных векторов) и нейронными сетями. По сравнению с ними деревья решений с градиентным бустином и линейные модели показывают худшие результаты точности детектирования адаптивных методов стеганографии в силу того, что они наиболее склонны к переобучению. Следует отметить, что метод *k*-ближайших соседей требует больших вычислительных затрат при классификации одного изображения, что неприменимо для реальных систем стеганоанализа. Метод опорных векторов требует определенного количества времени на обучение модели, так же как и методы на основе искусственных иммунных систем и нейросетей.

Список литературы

1. **Gulášová M., Jókay M.** Steganalysis of stegostorage library // Tatra Mountains Mathematical Publications. 2016. Vol. 67, № 1. P. 99–116. DOI 10.1515/tmmp-2016-0034
2. **Fridrich J.J., Goljan M., Hogea D.** Steganalysis of JPEG Images: Breaking the F5 Algorithm // 5th International Workshop on Information Hiding. 2002. DOI 10.1007/3-540-36415-3
3. **Hendrych J., Ličev L.** Advanced methods of detection of the steganography content // Lecture Notes in Electrical Engineering. 2020. Vol. 554. P. 484–493. DOI 10.1007/978-3-030-14907-9_47
4. **Yousfi Y. Butora J., Fridrich J., Giboulot Q.** Breaking Alaska: Color separation for steganalysis in JPEG domain // IH and MMSec 2019 - Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. 2019. P. 138–149. DOI 10.1145/3335203.3335727
5. **Saito T., Zhao Q., Naito H.** Second Level Steganalysis - Embeding Location Detection Using Machine Learning // 2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings. IEEE, 2019. P. 1–6. DOI 10.1109/ICAwST.2019.8923205.
6. **Butora J., Fridrich J.** Reverse JPEG Compatibility Attack // IEEE Transactions on Information Forensics and Security. IEEE, 2020. Vol. 15, № c. P. 1444–1454. DOI 10.1109/TIFS.2019.2940904
7. **Shniperov A. N., Prokofieva A. V.** Steganalysis Method of Static JPEG Images Based on Artificial Immune System // Automatic Control and Computer Sciences. 2020. Vol. 54, № 5. DOI 10.3103/S0146411620050077
8. **Yang Z. Wang K., Ma S., Huang Y., Kang X., Zhao X.** IStego100K: Large-scale Image Steganalysis Dataset // Digital Forensics and Watermarking. IWDW 2019. Lecture Notes in Computer Science. 2019. Vol. 12022. DOI 10.1007/978-3-030-43575-2_29

9. **Fridrich J., Pevný T., Kodovský J.** Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities // MM and Sec'07 - Proceedings of the Multimedia and Security Workshop 2007. 2007. DOI 10.1145/1288869.1288872
10. **Holub V., Fridrich J., Denemark T.** Universal distortion function for steganography in an arbitrary domain // Eurasip Journal on Information Security. 2014. Vol. 2014. DOI 10.1186/1687-417X-2014-1
11. **Guo L., Ni J., Su W., Tang C., Shi Y.** Using Statistical Image Model for JPEG Steganography: Uniform Embedding Revisited // IEEE Transactions on Information Forensics and Security. 2015. Vol. 10, № 12, DOI 10.1109/TIFS.2015.2473815
12. **Pevny T., Fridrich J.** Merging Markov and DCT features for multi-class JPEG steganalysis. 2007. P. 650503, DOI 10.1117/12.696774
13. **Vakhrushev A. et al.** LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. 2021. [Электронный ресурс] URL: https://www.researchgate.net/publication/354379217_LightAutoML_AutoML_Solution_for_a_Large_Financial_Services_Ecosystem (дата обращения: 01.11.2021).
14. **Дасгупта Д.** Искусственные иммунные системы и их применение / Под ред. Романюха А. ФИЗМАТЛИТ, 2006. 344 с.
15. **Pérez J.D.J.S., Rosales M.S., Cruz-Cortés N.** Universal steganography detector based on an artificial immune system for JPEG images // Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. 2017. Pp. 1896–1903, DOI 10.1109/TrustCom.2016.0290

References

1. **Gulášová M., Jókay M.** Steganalysis of stegostorage library. *Tatra Mountains Mathematical Publications*, 2016, vol. 67, no. 1, pp. 99–116. DOI 10.1515/tmmp-2016-0034
2. **Fridrich J. J., Goljan M., Hogea D.** Steganalysis of JPEG Images: Breaking the F5 Algorithm. 5th International Workshop on Information Hiding, 2002. DOI 10.1007/3-540-36415-3
3. **Hendrych J., Ličev L.** Advanced methods of detection of the steganography content. Lecture Notes in Electrical Engineering, 2020, vol. 554, pp. 484–493. DOI 10.1007/978-3-030-14907-9_47
4. **Yousfi Y. Butora J., Fridrich J., Giboulot Q.** Breaking Alaska: Color separation for steganalysis in JPEG domain. IH and MMSec 2019 - Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019, pp. 138–149. DOI 10.1145/3335203.3335727
5. **Saito T., Zhao Q., Naito H.** Second Level Steganalysis - Embeding Location Detection Using Machine Learning. 2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings. IEEE, 2019. Pp. 1–6. DOI 10.1109/ICAwST.2019.8923205
6. **Butora J., Fridrich J.** Reverse JPEG Compatibility Attack. IEEE Transactions on Information Forensics and Security. *IEEE*, 2020, vol. 15, no. c, pp. 1444–1454. DOI 10.1109/TIFS.2019.2940904
7. **Shniperov A. N., Prokofieva A. V.** Steganalysis Method of Static JPEG Images Based on Artificial Immune System. *Automatic Control and Computer Sciences*, 2020, vol. 54, no. 5. DOI 10.3103/S0146411620050077
8. **Yang Z. Wang K., Ma S., Huang Y., Kang X., Zhao X.** IStego100K: Large-scale Image Steganalysis Dataset. Digital Forensics and Watermarking. IWDW 2019. Lecture Notes in Computer Science, 2019, vol. 12022. DOI 10.1007/978-3-030-43575-2_29
9. **Fridrich J., Pevný T., Kodovský J.** Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities. MM and Sec'07 - Proceedings of the Multimedia and Security Workshop 2007, 2007. DOI 10.1145/1288869.1288872

10. **Holub V., Fridrich J., Denemark T.** Universal distortion function for steganography in an arbitrary domain. Eurasip Journal on Information Security. 2014, vol. 2014. DOI 10.1186/1687-417X-2014-1
11. **Guo L., Ni J., Su W., Tang C., Shi Y.** Using Statistical Image Model for JPEG Steganography: Uniform Embedding Revisited. IEEE Transactions on Information Forensics and Security, 2015, vol. 10, no. 12, DOI 10.1109/TIFS.2015.2473815
12. **Pevny T., Fridrich J.** Merging Markov and DCT features for multi-class JPEG steganalysis. 2007. P. 650503, doi: 10.1117/12.696774
13. **Vakhrushev A. et al.** LightAutoML: AutoML Solution for a Large Financial Services Ecosystem [Online]. 2021. URL: https://www.researchgate.net/publication/354379217_LightAutoML_AutoML_Solution_for_a_Large_Financial_Services_Ecosystem (01.11.2021).
14. **Dasgupta D.** Iskusstvennye immunnye sistemy i ih primenenie; Ed. A. Romanyuha. FIZMATLIT, 2006. 344 p. (in Russ.).
15. **Pérez J. D. J. S., Rosales M. S., Cruz-Cortés N.** Universal steganography detector based on an artificial immune system for JPEG images. Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. 2017. Pp. 1896–1903, DOI 10.1109/TrustCom.2016.0290

Информация об авторах

Александра Владимировна Прокофьева, аспирант кафедры прикладной математики и компьютерной безопасности

Алексей Николаевич Шниперов, кандидат технических наук, доцент кафедры прикладной математики и компьютерной безопасности

Information about the Authors

Aleksandra V. Prokofieva, postgraduate student of the Department of Applied Mathematics and Computer Security, Siberian Federal University

Alexey N. Shniperov, Candidate of Sciences in Technology, assistant Professor of the Department of Applied Mathematics and Computer Security, Siberian Federal University

Статья поступила в редакцию 27.11.2022;
одобрена после рецензирования 26.01.2023; принята к публикации 26.01.2023

The article was submitted 27.11.2022;
approved after reviewing 26.01.2023; accepted for publication 26.01.2023