Grzegorz GANCARCZYK [2], Agnieszka DĄBROWSKA – BORUCH [1, 2], Kazimierz WIATR [1, 2]
[1] AKADEMIA GÓRNICZO – HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE, al. Mickiewicza 30, 30-059 Kraków
[2] AKADEMICKIE CENTRUM KOMPUTEROWE CYFRONET AGH, ul. Nawojki 11, 30-950 Kraków

Efektywność parametrów statystycznych w detekcji informacji szyfrowanej

Mgr inż. Grzegorz GANCARCZYK

Absolwent kierunku Elektronika i Telekomunikacja (2009) oraz student V roku kierunku Elektrotechnika na Wydziale Elektrotechniki, Automatyki, Informatyki i Elektroniki AGH w Krakowie. Obecnie pracownik ACK Cyfronet AGH i członek tamtejszego Zespołu Akceleracji Obliczeń. Jego zainteresowania związane są z probabilistyką, procesami stochastycznymi, zjawiskiem szumu, cyfrowym przetwarzaniem sygnałów oraz akceleracją obliczeń numerycznych z wykorzystaniem logiki reprogramowalnej.

e-mail: g.gancarczyk@cyfronet.pl

Dr inż. Agnieszka DĄBROWSKA - BORUCH

Absolwentka kierunku Elektronika i Telekomunikacja na Wydziale Elektrotechniki, Automatyki, Informatyki i Elektroniki AGH (2002), dr nauk technicznych (2007). Obecnie jest adiunktem w Katedrze Elektroniki AGH oraz członkiem Zespołu Akceleracji Obliczeń ACK Cyfronet AGH. Jej zainteresowania naukowe to kompresja obrazu, systemy czasu rzeczywistego, układy programowalne oraz rekonfigurowalne. Prywatnie lubi sport, wycieczki krajoznawcze, dobrą muzykę.

e-mail: adabrow@agh.edu.pl

Prof. dr hab. inż. Kazimierz WIATR

Studia AGH Kraków (1980), doktor nauk technicznych (1987), doktor habilitowany (1999) i profesor (2002). Profesor zwyczajny w Akademii Górniczo - Hutniczej oraz dyrektor Akademickiego Centrum Komputerowego Cyfronet AGH. Prowadzone prace badawcze dotyczą komputerowego sterowania procesami, systemów wizyjnych, systemów wieloprocesorowych, układów programowalnych, rekonfigurowalnych systemów obliczeniowych i sprzętowych metod akceleracji obliczeń.

e-mail: wiatr@agh.edu.pl

Streszczenie

Informacja szyfrowana, podobnie jak wszystkie inne typy danych, może zostać poddana analizie statystycznej. Wyznaczenie dla niej parametrów takich jak wartość średnia, wariancja czy też entropia nie nastręcza większych trudności. Wykorzystać do tego można nowoczesne narzędzia numeryczne jak np. MATLAB, Mathcad czy też Microsoft Exel. Pytanie, na które ma dać odpowiedź niniejsze opracowanie brzmi – *„czy parametry te niosą ze sobą wiedzę, którą można wykorzystać w użyteczny sposób?"* Przykładowym zastosowaniem może być np. określenie czy informacja jest zaszyfrowana (ang. *cipher text*), czy też jest ona jawna (ang. *plain text*).

Słowa kluczowe: szyfrowanie, parametry statystyczne, analiza danych, rozkład statystyczny.

Effectiveness of statistic parameters in cipher data detection

Abstract

A cipher text, like any other data, can be analysed with use of parameters typical for statistics. Values such as the mean value, variance or entropy are easy to be calculated, especially if one can use numerical tools like e.g. MATLAB, Mathcad or simply Microsoft Exel. The question, to which this paper should give an answer is – *"do those parameters provide any information that could be used in any useful way?"* For example, the information, whether the analysed data is a cipher or plain text. The available publications about distinguishing the cipher from plain text use only methods typical for testing the randomness of cipher text and random number generator or immunity for cipher breaking. They are presented in the paper by the National Institute of Standards and Technology [1]. The other common method, used for distinguishing the data, is the analysis based on entropy [2]. Lack of published results about the efficiency of methods based on e.g. entropy, is additional motivation for this paper. (see Paragraph 1.) The proposed algorithms use parameters and transformations typical for Statistic and Signal Processing to classify the analysed data as cipher/plain. The authors assume that cipher data are very similar to random numbers due to Shannon's *Perfect Secrecy* theorem [3]. Six types of plain and cipher data (text, music, image, video, archives and others), seven types of cipher cores (3DES, AES, Blowfish, CAST – 128, RC4, Serpent, Twofish) and various length (1 B to 2323 B) data were examined and group of the so called Statistic Parameters was formed (see Table 1). Definitions of all of them (and a few more) are given by equations (1) to (12). The efficiency of Statistic Parameters after 1417 test samples is shown in Table 2. The most interesting results are also shown in Figs. 1 to 9. (see Paragraphs 2 – 4.) The results show that using simple values like e.g. energy one can built a data distinguisher of the efficiency equal to 90% and low numerical complexity. The lower bound for usability of this method was found to be 200 B. The upper bound was not found. The presented algorithm can be used for creating a network data analyser or cipher text detector. (see Paragraph 5.)

Keywords: cipher, cryptography, cryptanalysis, statistic parameters, data analysis, probability distribution, noise.

1. Wstęp

Kryptografia stanowi jedną z ważniejszych części nowoczesnej informatyki. Świadczy o tym fakt, że bardzo często wymieniana jest jako odrębna dziedzina wiedzy. Pojawia się również jako część składowa nauk informatyczno – technicznych.

Nowoczesna kryptologia to, prócz metod i sposobów zabezpieczania informacji przed niepowołanym dostępem, również kryptoanaliza, czyli nauka o łamaniu zabezpieczeń kryptograficznych. Właśnie na potrzeby tej drugiej przeprowadzono badania, mające na celu określenie przydatności podstawowych metod statystycznej analizy danych i cyfrowego przetwarzania danych do klasyfikacji informacji cyfrowej jako zaszyfrowanej/niezaszyfrowanej.

Statystyczna analiza danych i przetwarzanie sygnałów definiują bardzo dużą liczbę parametrów pomocnych w identyfikacji procesów i sygnałów. Ich liczba jest, co prawda skończona, lecz niewątpliwie bardzo duża. Ze względu na ograniczony czas i środki, niemożliwym było przebadanie ich wszystkim pod kątem przydatności do detekcji informacji niejawnej. Pierwszą grupę alternatywnych dla zaprezentowanych w niniejszym artykule metod, stanowią wielkości i wytyczne zdefiniowane przez National Institute of Standards and Technology w [1]. Przykładowo są to testy: częstotliwości, rzędu macierzy binarnych, Maurera. Wymienionym w [1], ale wartym osobnego wspomnienia parametrem jest entropia. Ciekawą wydaje się być zwłaszcza rodzima publikacja dotycząca wykorzystania jej, nie tylko na potrzeby kryptoanalizy, ale również np. steganografii [2]. Brak kompleksowych publikacji dotyczących detekcji informacji szyfrowanej (od pomysłu, aż do ostatecznej realizacji i testów) był dodatkową motywacją do przedstawienia wyników prac w niniejszym artykule.

2. Wprowadzenie

W roku 1949 Claude E. Shannon, w swojej pracy *"Communication Theory of Secrecy System",* zdefiniował dokładnie pojęcie idealnej tajności (ang. *The Perfect Secrecy*) [3]. Według niego, warunkiem otrzymania na wyjściu systemu szyfrującego idealnie tajnej wiadomości jest użycie klucza jednorazowego (ang. *one – time pad key*). Klucz ten oprócz ograniczenia, że może być użyty tylko raz, dodatkowo charakteryzuje się długością nie mniejszą od długości utajnianej wiadomości oraz pełną przypadkowością występujących w nim symboli. Występujące w kluczu wartości posiadają

więc płaski rozkład gęstości prawdopodobieństwa. Przy spełnieniu tych warunków, dla danych wyjściowych systemu szyfrującego pojawia się następująca właściwość – *żaden z symboli występujących w informacji zaszyfrowanej nie jest bardziej prawdopodobny niż inne* [4].

Stosowanie kluczy typu *one – time pad* jest w rzeczywistości niemożliwe ze względu na brak generatorów liczb losowych oraz konieczność generowania klucza o długości równej długości wiadomości jawnej [4, 5, 6, 7]. Pewnym sposobem na poradzenie sobie z tymi problemami jest stosowanie pseudolosowych kluczy o ograniczonej długości (typowo od 128 do 1024 bitów dla kluczy symetrycznych oraz od 1024 do 4096 dla kluczy asymetrycznych). Zabieg ten prowadzi do lepszego lub gorszego spełnienia warunku *confusion and diffusion* zdefiniowanego również przez Shannona w [3]. *Confusion* ma doprowadzić do jak najbardziej przypadkowej relacji pomiędzy kluczem, a wiadomością zaszyfrowaną (jak najmniejsza ich wzajemna korelacja osiągana przez losową naturę klucza oraz wiadomości zaszyfrowanej). Operacje wchodzące w skład *diffusion* „wygładzają" natomiast rozkład danych wyjściowych. Innymi słowy czynią go bardziej płaskim.

Wspomnieć należy jeszcze o trzech najczęściej występujących topologiach systemów szyfrujących i przykładowych rozwiązaniach, które je wykorzystują. Są to odpowiednio: sieci Feistela (np. 3DES), sieci substytucyjno – permutacyjne (np. AES), sieci oparte o proste operacje logiczne (np. RC4).

Pierwsze dwie wykorzystywane są w blokowych systemach szyfrujących. Ostatnia znajduje zastosowanie głównie w strumieniowych algorytmach szyfrowania danych.

Podział wiadomości na bloki pozwala ograniczyć długość koniecznego do wygenerowania klucza, jednakże towarzyszące mu przekształcenia są znacznie trudniejsze w wykonaniu niż operacje logiczne na strumieniu danych i kluczu. Podobna sytuacja występuje w przypadku zastosowania wyrafinowanych bloków substytucyjnych (ang. *substitution – boxes*), permutacyjnych (ang. *per-mutation – boxes*) i rozszerzeń (ang. *expansion – boxes*).

Mając to wszystko na uwadze można napisać, że stosowane obecnie algorytmy szyfrowania oferują niemalże idealną tajność informacji.

Od tego miejsca przyjmuje się, że dane zaszyfrowane pochodzące z systemu szyfrującego charakteryzują się rozkładem bardzo zbliżonym do równomiernego (płaskiego), a w przypadku idealnym, powinny posiadać rozkład płaski. Można więc interpretować je jako próbki cyfrowego szumu białego (przez „cyfrowy" należy rozumieć „o wartościach dyskretnych").

3. Parametry statystyczne

3.1. Dane

Ponieważ w rozważaniach ujęte zostają jedynie systemy szyfrujące operujące na danych binarnych w oparciu o najpopularniejsze spośród stosowanych obecnie algorytmów szyfrujących, toteż zasadnym wydaje się wspomnieć o sposobie interpretacji danych jawnych, klucza oraz danych zaszyfrowanych.

Istnieje wiele sposobów interpretacji posiadanej informacji cyfrowej (tekst, dźwięk, obraz, itp.), tak samo jak istnieje wiele „kwantów" danych (bit, bajt, 16 bitów, 24 bity, 32 bity, itd.). Poszczególne „kwanty" znajdują zastosowanie na różnych poziomach abstrakcji (bit – logika niskiego poziomu, 64 bity – typ double języka programowania). Wiele z nich wykorzystuje się do opisu tego samego (24 bity, 32 bity i 48 bitów – trzy najpopularniejsze sposoby opisu ścieżki dźwiękowej). System binarny niejako narzuca konieczność operowania na „kwancie", będącym wielokrotnością liczby 2.

Uznano, że optymalną jednostką w tych rozważaniach będzie bajt, równy ośmiu bitom, interpretowanym jako liczba całkowita bez znaku z przedziału wartości od 0 do 255. Równie poprawną jednostką mogłyby być 2 lub 4 bity, czy też sam pojedynczy bit, jednakże wykonywanie operacji matematycznych i logicznych charakterystycznych dla statystycznej analizy danych i cyfrowego przetwarzania danych, byłoby znacznie wolniejsze, a przez to mniej wydajne (praktycznie nie spotyka się jednostek arytmetyczno – logicznych z magistralą mniejszą niż 8 bitów) [8, 9]. Odrzucenie „kwantów" większych niż 8 bitów wynika z prostej przyczyny – jak podzielić, a następnie przetworzyć informację z pliku o długości równej np. 10 B? W przypadku „kwantu" równego 32 bity (tj. 4 B) w wyniku podzielenia omawianej informacji na części o wielkości jednego „kwantu", otrzymanych zostanie 2,5 „kwantu". Jak więc należy zinterpretować to 0,5 „kwantu" zważywszy, że powinien być on niepodzielny? Prócz tego należy pamiętać, że najczęściej szyfrowanymi danymi jest tekst. W jego przypadku poszczególne litery i znaki reprezentowane są jako liczby znajdujące odwzorowanie w 8 bitowej tablicy ASCII.

Przebadane algorytmy szyfrujące, to: 3DES, AES, Blowfish, CAST – 128, RC4, Serpent, Twofish.

W grupie tej znajdują się zarówno przedstawiciele systemów szyfrujących o topologii blokowej, jak i strumieniowej. Inny podział, który można również zastosować, to podział na algorytmy często stosowane w rozwiązaniach informatycznych, np. protokół SSL (3DES, AES), jak również rzadziej występujące (Serpent, Twofish).

Jako danymi wejściowymi (jawnymi) posłużono się grupą 25, bardzo często spotykanych formatów zapisu informacji. Grupę tę podzielić można na 6 następujących podgrup: dane tekstowe (pliki *.txt ANSI, *.txt UTF – 8, *.txt Unicode, *.txt Big Endian i *.rtf), dane muzyczne (pliki *.aiff, *.mid, *.mp3, *.wav i *.wma), dane graficzne (pliki *.bmp, *.gif, *.ico, *.jpg i *.png), dane wideo (pliki *.asf, *.avi i *.mov), dane zarchiwizowane (pliki *.cab, *.jar, *.rar, *.tar i *.zip), dane różne (pliki *.html i *.exe).

Choć wydawać się może, iż pliki z rozszerzeniem .html zaliczyć należy do podgrupy danych tekstowych, to jednak występujące w nich znaki specjalne i zwroty są charakterystyczne jedynie dla języków programowania. Konsekwencją tego jest to, że wartości parametrów statystycznych wyznaczonych dla tych plików znacząco odbiegają od wartości tych samych parametrów wyznaczonych dla podgrupy dane tekstowe.

Do przeprowadzenia analizy użyto plików o długości od 1 B do 2323 B (typowo od 200 B do 1500 B).

Zarówno do wykonania analiz jak i wygenerowania wyników posłużono się narzędziem numerycznym MATLAB firmy Mathworks oraz konwerterem HEX to DEC [10].

3.2. Parametry statystyczne

Analizę statystyczną danych przeprowadzono w oparciu o następujące parametry i przekształcenia: wartość średnia, energia, moc średnia, wartość skuteczna, wariancja, odchylenie standardowe, momenty normalne od zerowego do piątego, unormowane momenty zwykłe, momenty centralne od zerowego do piątego, unormowane momenty centralne, odcięta środka ciężkości kwadratu, wariancja wokół odciętej środka ciężkości kwadratu, szerokość średniokwadratowa, transformata Fouriera, moc średnia widmowa, histogram, autokorelacja, autokowariancja.
Z grupy tej jako przydatne do określenia czy dane są zaszyfrowane, wybrano jedynie niektóre. Parametry te zestawiono w tabeli 1 i, od tej pory, grupę tych czternastu wielkości będzie się nazywać w skrócie Parametrami Statystycznymi. Dobór Parametrów Statystycznych polegał na wykluczeniu tych, spośród wymienionych wcześniej, dla których wyniki nie posiadały żadnych cech charakterystycznych, dających się zauważyć w przypadku analizy graficznej (wykresy) jak i numerycznej (wartości wyników). Cechy te to np. skupienie wyników wokół pewnej wartości stałej (np. wariancja) lub funkcji (np. energia), czy też charakterystyczny kształt wykresu (np. histogram). Analizę przeprowadzono na ograniczonej liczbie próbek danych jawnych i zaszyfrowanych.

Przyglądając się parametrom z tabeli 1, zauważyć można obecność takich wielkości jak energia, moc średnia, czy też moc średnia widmowa (wartość skuteczna jest często spotykana w statystyce pod nazwą wartości średniokwadratowej). Wielkości te są zazwyczaj związane z cyfrowym przetwarzaniem sygnałów

i charakterystyczne dla sygnałów niosących ze sobą informację. Jak wspomniano wcześniej, dane rozpatrywać można również jako próbki n–wymiarowego sygnału (szumu), dlatego też wielkości te zostały włączone do grupy Parametrów Statystycznych. Nie podlega dyskusji, iż zarówno dane jawne jak i zaszyfrowane niosą ze sobą informacje.

Tab. 1. Parametry Statystyczne przydatne do detekcji informacji zaszyfrowanej
Tab. 1. Statistic parameters useful for cipher detection

Zweryfikowane pozytywnie Parametry Statystyczne
Wartość średnia zmodyfikowana
Energia
Moc średnia
Wartość skuteczna
Moc średnia widmowa
Wariancja
Wariancja zmodyfikowana
Różnica wariancji
Odchylenie standardowe
Odchylenie standardowe zmodyfikowane
Różnica odchyleń standardowych
Momenty zwykłe
Momenty centralne
Histogram zmodyfikowany

Na koniec warto jeszcze wyjaśnić znaczenie przymiotnika „zmodyfikowany", pojawiającego się przy niektórych pozycjach z tabeli 1. Użycie parametru w brzmieniu identycznym jak jego definicja matematyczna nie zawsze przynosiło pożądany efekt. Przeprowadzenie pewnej modyfikacji jego definicji znacząco poprawiało efektywność detekcji zaszyfrowania danych. Stąd też obecność tego zwrotu dla odróżnienia definicji oryginalnej od definicji zaproponowanej przez autorów.

3.3. Definicje matematyczne

Poniżej przedstawiono definicje matematyczne wymienionych w tabeli 1 Parametrów Statystycznych. Definicje niezmodyfikowane zaczerpnięto z pozycji [9].

Wartość średnia funkcji dyskretnej dana jest wyrażeniem

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{n=N-1} x(n), \tag{1}$$

gdzie N jest ilością wszystkich próbek, a $x(n)$ jest wartością próbki o numerze n.

W procesie detekcji wykorzystywany jest parametr nazywany wartością średnią zmodyfikowaną. Składa się on z dwóch kroków. W pierwszym obliczana jest wartość średnia dla całości próbek zgodnie z definicją (1). Jeśli znak błędu obliczonego w stosunku do wartości wzorcowej wynoszącej 127,5 (wartość średnia funkcji dyskretnej o rozkładzie równomiernym, przyjmującej wartości spośród zbioru liczb całkowitych od 0 do 255) jest dodatni, dane zostają zakwalifikowane jako zaszyfrowane. Jeśli błąd jest ujemny i mniejszy niż -2%, dane zostają uznane za nieszyfrowane. W przeciwnym wypadku konieczne jest wykonanie drugiego etapu analizy w oparciu o wartość średnią w oknie. Obliczone zostaje wyrażenie

$$\bar{x}_W = \frac{1}{W}\sum_{n=0}^{n=W-1} x(n), \tag{2}$$

gdzie W jest rozmiarem okna. Rozmiar okna jest zawsze wielokrotnością liczby 2 i nie może być większy niż długość danych. Po obliczeniu wartości według (2), okno ulega przesunięciu o 1 (granice sumowania dla (2) wynoszą wtedy odpowiednio $n = 1$ i $n = W$), zostaje obliczona nowa wartość średnia w oknie. Cały proces jest powtarzany do momentu, gdy okno osiągnie koniec (górna granica sumowania równa indeksowi ostatniej próbki danych), wtedy następuje obliczenie wartości średniej ze wszystkich wcześniej otrzymanych wartości średnich dla danego okna W. Operację obliczenia wartości średniej w oknach przeprowadza się dla wszystkich możliwych wielkości okna. Dla każdego z przypadków zostaje obliczona wartość błędu. Jeśli w choć jednym przypadku znak błędu jest dodatni, dane zostają uznane za zaszyfrowane. W przeciwnym wypadku uważa się je za jawne.

Energię definiuje się jako

$$E_x = \sum_{n=0}^{n=N-1} x^2(n). \tag{3}$$

Moc średnia dana jest zależnością

$$\overline{P}_x = \frac{1}{N} E_x. \tag{4}$$

Definicja matematyczna wartości skutecznej funkcji dyskretnej ma postać

$$RMS_x = \sqrt{\overline{P}}. \tag{5}$$

Wariancję danych oblicza się przy użyciu zależności

$$\sigma_x^2 = \frac{1}{N}\sum_{n=0}^{n=N-1}\left[x(n) - \bar{x}\right]^2. \tag{6}$$

Możliwość wykonania modyfikacji parametru (6) wynika z dwóch sposobów interpretacji występującej w nim wartości średniej. Jako wartość średnią występującą w wariancji niezmodyfikowanej wykorzystuje się wzór (1). W przypadku wariancji zmodyfikowanej użyta zostaje wartość średnia teoretyczna równa 127,5. Definicję wariancji zmodyfikowanej można zapisać następująco

$$\sigma_{xM}^2 = \frac{1}{N}\sum_{n=0}^{n=N-1}\left[x(n) - 127{,}5\right]^2. \tag{7}$$

Różnica wariancji jest równa różnicy pomiędzy zależnościami (6) i (7).

Odchylenie standardowe dane jest zależnością jak poniżej

$$\sigma_x = \sqrt{\sigma_x^2}. \tag{8}$$

Identycznie jak dla wariancji, tutaj również można dokonać modyfikacji w oparciu o parametr wartości średniej i użyć wartości średniej własnej danych lub też wartości średniej teoretycznej. Ten drugi przypadek dany jest zależnością (9).

$$\sigma_{xM} = \sqrt{\sigma_{xM}^2}. \tag{9}$$

Różnica odchyleń standardowych jest zdefiniowana jako różnica wyrażeń danych zależnościami (8) i (9).

Momenty normalne dane są jako

$$\bar{k}_x^m = \sum_{n=1}^{n=N} n^m x(n), \tag{10}$$

gdzie m jest rzędem momentu.

W rozważaniach ograniczono się jedynie do pierwszych sześciu momentów, począwszy od zerowego, skończywszy na piątym. W stosunku do zależności od (1) do (9) następuje zmiana granic sumowania. Powodem jest niemożliwość pominięcia w trakcie

wykonywania operacji matematycznych jakiejkolwiek próbki. Sumowanie od 0 spowoduje, że pierwsza z próbek nie będzie miała żadnego wkładu do wyniku końcowego.

Momenty centralne dane są zależnością

$$\delta_x^m = \sum_{n=1}^{n=N} \left[n - \bar{k}_x^1\right]^m x(n) \cdot \tag{11}$$

Zmodyfikowany histogram otrzymuje się poprzez obliczenie wartości stosunku pomiędzy ilością próbek w pierwszym i ostatnim przedziale histogramu. Operację tą wykonuje się dla przedziału histogramu o szerokości równej 64. Jeśli stosunek jest mniejszy niż 1,25, wtedy dane zostają zakwalifikowane jako zaszyfrowane. W przeciwnym przypadku zostaje obliczona wartość stosunku dla histogramu o szerokości przedziału równej 128. Jeśli stosunek znajduje się w przedziale od 0,75 do 1,25, dane zostają uznane za zaszyfrowane. Jeśli jest to wartość z przedziału do 1,25 do 2,00, wtedy bada się trend zmian wartości stosunków dla histogramu o długości przedziału 64 i 128. Gdy jest on malejący (wartość stosunku dla przedziału o szerokości 128 mniejsza niż dla przedziału o szerokości 64), dane zostają uznane za zaszyfrowane. W pozostałych przypadkach (trend rosnący, wartość stosunku dla przedziału równego 128 większa od 2,00 lub mniejsza od 1,25) dane uznaje się za jawne. Zaproponowane szerokości przedziałów histogramu oraz graniczne wartości stosunków wynikają z przeprowadzonych testów. Przy zastosowaniu takich, a nie innych, założeń uzyskuje się najlepsze właściwości detekcyjne metody. Teoretyczny stosunek ilości elementów w przedziałach wynosi 1 i wynika z równomiernego rozkładu, który założono na wstępie.

Moc średnia widmowa zdefiniowana jest wzorem

$$P_{xx}\left(e^{j\omega}\right) = \frac{1}{2\pi(N-1)} \left| \sum_{n=0}^{n=N-1} x(n) e^{-j\omega n} \right| . \tag{12}$$

4. Wyniki

4.1. Informacje ogólne

Wyniki, obrazujące efektywność Parametrów Statystycznych w detekcji informacji zaszyfrowanej, przedstawiono na dwa sposoby. Dla wszystkich parametrów wymienionych w tabeli 1 przygotowano zestawienie i przedstawiono je w formie tabelarycznej (tabela 2). Wybrane przez autorów parametry przedstawiono również na rysunkach. Graficznie zilustrowano po trzy przykłady dla trzech różnych parametrów. Przykład pierwszy (ogólny) przedstawia położenie wartości Parametrów Statystycznych dla 1417 prób (na danych zarówno jawnych jak i zaszyfrowanych) względem wykresu funkcji teoretycznej (wzorcowej). Funkcja wzorcowa obrazuje zależność danego Parametru Statystycznego od długości danych wejściowych. Przykład drugi odnosi się do znacznie mniejszej liczby przypadków, o długości od 500 B do 750 B. Przypadek trzeci obrazuje największą niepewność uzyskanego wyniku – dane krótkie. Podobnie jak z każdymi innymi obliczeniami statystycznymi, tutaj również trafność, sprawność i sensowność stosowania metody ograniczone są przez liczbę dostępnych danych.

Proces weryfikacji przydatności Parametrów Statystycznych i określenia wartości progów tolerancji w celu maksymalizacji ich skuteczności, przeprowadzono na niewielkiej liczbie próbek. Liczba ta była kilkudziesięciokrotnie mniejsza od liczby próbek, którymi posłużono się do określenie efektywności przedstawionych w tabeli 2. Również różnorodność danych, którymi posłużono się na etapie weryfikacji była niewielka zarówno pod względem ilości typów danych jawnych (4 typy) jak i algorytmów wykorzystanych do zaszyfrowania danych (3 rodzaje).

Efektywność detekcji informacji zaszyfrowanej zostaje obliczona przy pomocy zależności

$$\eta_{\%} = \frac{100}{K} \sum_{i=1}^{i=K} \delta_i , \tag{13}$$

gdzie

$$\delta_i = \frac{c_i}{c_i + u_i} , \tag{14}$$

a K jest ilością składowych (m.in. .rtf, .avi, itd.), δ_i jest prawdopodobieństwem prawidłowego rozróżnienia i-tego typu plików, c_i to ilość wszystkich poprawnie zakwalifikowanych i-tych plików, a u_i to ilość wszystkich niepoprawnie zakwalifikowanych i-tych plików.

Przyglądając się wyrażeniom (13) i (14) łatwo zauważyć, że typ pliku, liczba plików danego typu oraz sposób ich szyfrowania nie ma żadnego wpływu na wynik końcowy efektywności.

4.2. Efektywność parametrów statystycznych

Połowa wymienionych w tabeli 2 Parametrów Statystycznych wykorzystywanych do detekcji informacji zaszyfrowanej, posiada efektywność większą niż 90%. Na szczególną uwagę zasługują zwłaszcza cztery – energia, różnica wariancji, momenty centralne oraz histogram zmodyfikowany. Pozostałe, choć ich efektywność przekracza 90%, są w większości wielkościami pochodnymi lub bezpośrednio związanymi z wcześniej wymienionymi. Spośród nich tylko energia nie jest związana bezpośrednio ze statystyczną analizą danych [8].

Tab. 2. Efektywność Parametrów Statystycznych w detekcji szyfrowania informacji
Tab. 2. Effectiveness of statistic parameters in cipher information detection

Parametry Statystyczne	Efektywność [%]
Wartość średnia zmodyfikowana	80,61
Energia	91,51
Moc średnia	90,26
Wartość skuteczna	90,26
Wariancja	88,67
Odchylenie standardowe	87,27
Wariancja zmodyfikowana	89,06
Odchylenie standardowe zmodyfikowane	89,19
Różnica wariancji	92,20
Różnica odchyleń standardowych	91,91
Momenty normalne od zerowego do piątego	89,91
Momenty centralne od zerowego do piątego	91,59
Histogram zmodyfikowany	91,17
Moc średnia widmowa	89,62

Dwie z wymienionych metod – energia oraz histogram zmodyfikowany – wykazują bardzo dobrą efektywność w poprawnej klasyfikacji danych nieszyfrowanych. Momenty centralne z kolei charakteryzują się najlepszą efektywnością poprawnej klasyfikacji danych zaszyfrowanych. Parametr jakim jest różnica wariancji, choć nie tak dobry w odrębnej klasyfikacji danych, globalnie myli się w najmniejszym stopniu jeśli chodzi o rozróżnianie typu danych.

Wspomniane wcześniej: energia i histogram zmodyfikowany, wymagają wykonania najmniejszej liczby operacji matematycznych i przekształceń, spośród wszystkich czterech wymienionych metod. Ponadto nie wymagają wykonania żadnych dodatkowych obliczeń i operacji, tak jak ma to miejsce w przypadku momentów centralnych i różnicy wariancji. W ich przypadku, koniecznym staje się obliczenie momentu zwykłego pierwszego rzędu oraz wykonanie obliczeń dla sześciu rzędów parametru (momenty centralne) lub policzenia dwóch różnych wariancji (różnica wariancji).

4.3. Interpretacja graficzna

W niniejszym paragrafie przedstawiono, w sposób graficzny, rozkład wartości energii, pierwszego momentu centralnego oraz różnicy odchyleń standardowych w funkcji długości informacji testowej. W sposób wyraźny odróżniono przypadki odpowiadające informacji jawnej (symbol ◊) od przykładowych próbek informacji zakodowanej (symbol □). Na każdej z ilustracji pojawia się również linia, obrazująca przebieg funkcji wzorcowej.

Dla każdego z Parametrów wykonano trzy wykresy obrazujące – wszystkie przypadki testowe, próbki o długości z wybranego przedziału, dane o niewielkiej długości.

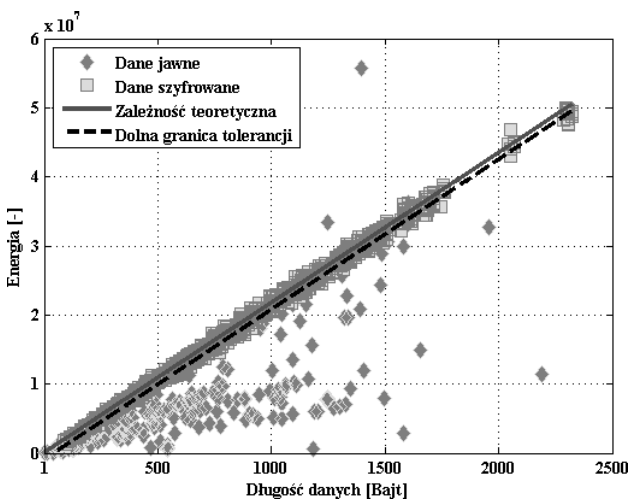
4.3.1. Energia

Algorytm klasyfikacji danych jako zaszyfrowane/jawne działa w następujący sposób:

- wyznaczona zostaje wartość teoretyczna dla Parametru energia w oparciu o funkcję $f(N) = 21717{,}5 \cdot N$, gdzie N to długość danych, a $f(N)$ to teoretyczna wartość energii danych (zależność wyznaczono w oparciu o równanie (3) dla procesu losowego o rozkładzie płaskim i wartościach z przedziału od 0 do 255),
- wyznaczona zostaje wartość dolnej granicy tolerancji dla Parametru energia w oparciu o równanie $g(N) = 21717{,}5 \cdot (N-42)$),
- na podstawie wszystkich bajtów danych, obliczona zostaje ich energia (według zależności (3)),
- przypadki o energii mniejszej niż dolna granica tolerancji, zakwalifikowane zostają jako dane jawne, zaś pozostałe jako dane zaszyfrowane.

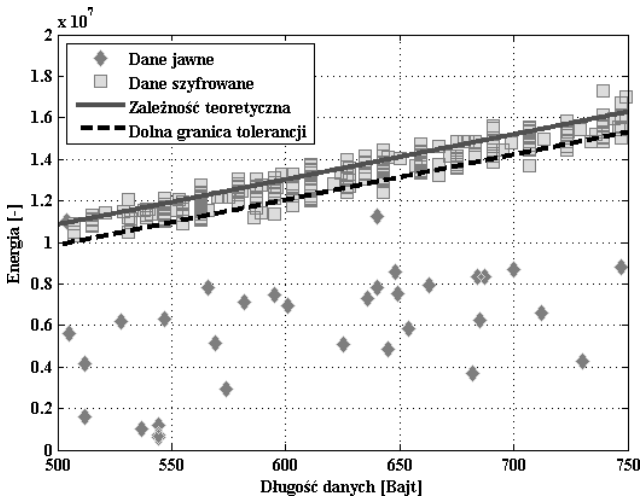
W tym miejscu należy się jeszcze słowo wyjaśnienia na temat tego jak dokonano wyboru równania opisującego przebieg prostej będącej dolną granicą tolerancji – dokonano tego w sposób doświadczalny. Operując na wybranej liczbie danych testowych wybrano takie równanie prostej, dla którego liczba niepoprawnie zakwalifikowanych danych była najmniejsza.

Na rysunku 1 skupienie danych wokół prostej teoretycznej jest bardzo dobrze widoczne. Kolejna zauważalna prawidłowość, to zgodne z oczekiwaniami, rosnące wraz z długością odchylenie danych jawnych od krzywej teoretycznej. Dokładnie odwrotny proces zachodzi w przypadku danych zaszyfrowanych.



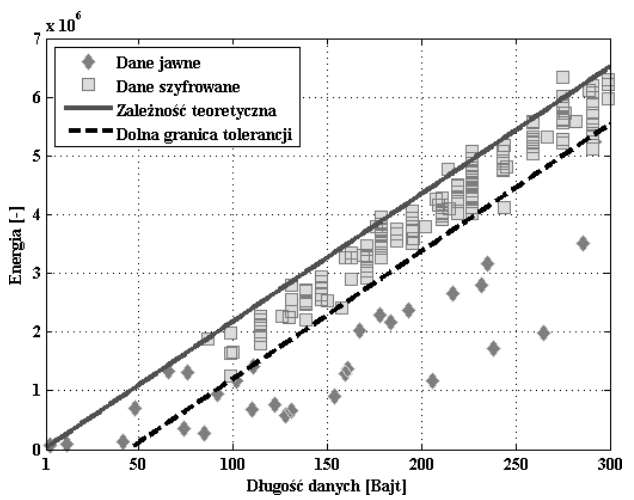
Rys. 1. Zależność energii danych w funkcji ich długości
Fig. 1. Relation between energy and length of data

Przyglądając się wykresowi z rysunku 2 zauważyć można, że błąd popełniany w przypadku danych jawnych występuje jedynie raz (okolice 500 B). Jest on znacznie częstszy dla danych zaszyfrowanych. Widoczne jest również rosnące odchylenie energii danych nieszyfrowanych od wartości teoretycznej, wraz ze wzrostem długości tych danych. W przypadku danych zaszyfrowanych, niezmiennie są one skupione wokół prostej teoretycznej.



Rys. 2. Zależność energii danych w funkcji ich długości dla wybranego przedziału długości
Fig. 2. Relation between energy and length of data for chosen data length

Doskonale widocznym na rysunku 3 jest, że równanie opisujące prostą będącą dolną granicą tolerancji, całkowicie uniemożliwia poprawną klasyfikację danych krótszych niż około 50 B. Energia musiałaby być ujemna, aby klasyfikacja mogła być poprawna. Taki przypadek zgodnie z definicją (3) nigdy nie zaistnieje, przy operowaniu wartościami ze zbioru liczb całkowitych.



Rys. 3. Zależność energii danych w funkcji ich długości dla danych krótkich
Fig. 3. Relation between energy and length of data for short data

Granica sensu stosowalności metody występuje dla danych o długości z przedziału od 150 B do 200 B. Dopiero w nim zauważyć można widoczne odchylenie wartości energii danych jawnych od dolnej granicy tolerancji. Wartość tej odchyłki wzrasta wraz ze wzrostem długości danych. Bezpiecznie zatem przyjąć, że parametr energia może zostać zastosowany w metodzie detekcji szyfrowania informacji w przypadku, gdy ma się do czynienia z informacją dłuższą niż 200 B.

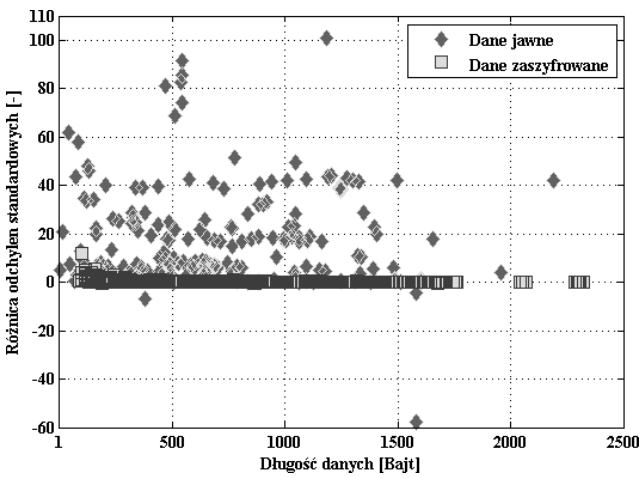
4.3.2. Różnica odchyleń standardowych

Algorytm klasyfikacji danych jako zaszyfrowane/jawne działa w następujący sposób:

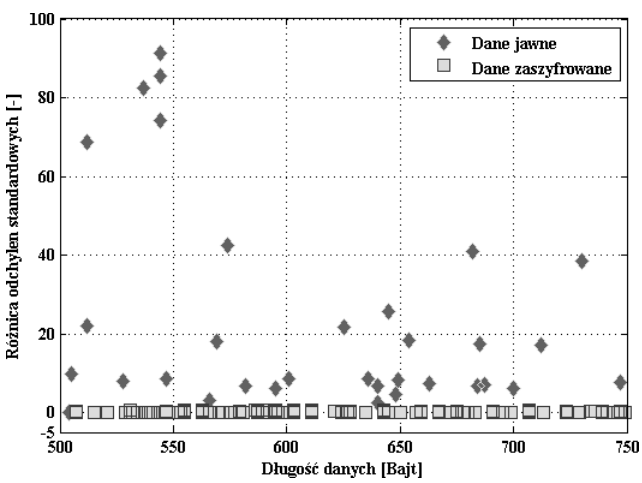
- na podstawie wszystkich bajtów danych obliczone zostają wartości ich odchylenia standardowego oraz odchylenia standardowego zmodyfikowanego (według (8) i (9)),

- przypadki, dla których wartości różnicy odchyleń standardowych nie mieszczą się wewnątrz przedziału skupionego wokół wartości 0 (przedział zawiera się w granicach od -1,20 do 2,25 i został wyznaczonych doświadczalnie w celu maksymalizacji sprawności metody), zakwalifikowane zostają jako dane jawne, pozostałe natomiast jako dane zaszyfrowane.

Na rysunku 4 przedstawiono przypadek ogólny dla wszystkich przygotowanych próbek danych. W przypadku danych zaszyfrowanych, których teoretyczna wartość średnia powinna być na poziomie 127,5, różnica odchyleń standardowych powinna wynieść 0. Przyglądając się umiejscowieniu punktów odpowiadającym informacji tajnej zauważyć można, że trend taki występuje. Podobnie jak poprzednio, wraz ze wzrostem długości danych szyfrowanych wzrasta ich skupienie wokół wartości teoretycznej. W przypadku danych jawnych, proces zwiększania się odchyłki od wartości teoretycznej nie jest dobrze widoczny, jednakże odchyłki te są na tyle odbiegające od dolnego i górnego progu tolerancji, że możliwa jest ich poprawna klasyfikacja. Zgodnie z danymi z tabeli 2, prawidłowa klasyfikacja typu danych jest dla tego Parametru najlepsza spośród wszystkich zilustrowanych w niniejszym podrozdziale.



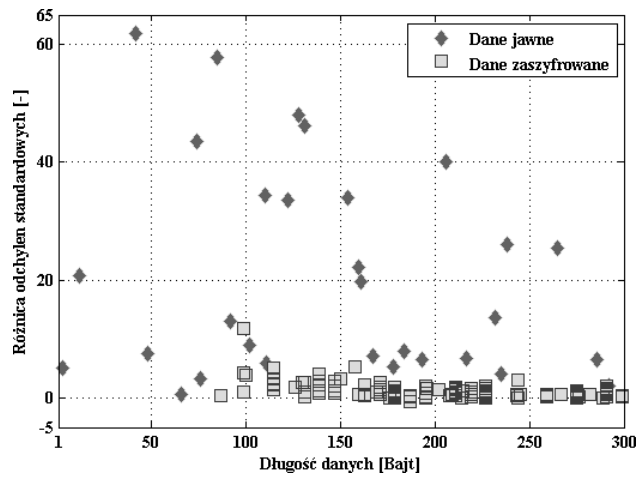
Rys. 4. Zależność różnicy odchyleń standardowych danych w funkcji ich długości
Fig. 4. Relation between difference of standard deviations and length of data



Rys. 5. Zależność różnicy odchyleń standardowych danych w funkcji ich długości dla wybranego przedziału długości
Fig. 5. Relation between difference of standard deviations and length of data for chosen data length

Na rysunku 5, pomimo wykonania zbliżenia na dane o określonej długości, brak jest widocznego wzrostu odchyłki od wartości teoretycznej wraz ze wzrostem długości danych. Zauważyć jednak można praktycznie stałe wartości odchyleń dla danych zaszyfrowanych w całym analizowanym przedziale długości danych.

Widoczne na rysunku 6 znaczne odchyłki różnicy odchyleń standardowych danych zaszyfrowanych sprawiają, że zasadność stosowania metody można rozważać dopiero dla danych dłuższych niż 200 B, a nawet 250 B. Co warte zauważenia, dane jawne wykazują znaczne odchyłki dla całego analizowanego przedziału długości. Można stwierdzić, że metoda bardzo dobrze radzi sobie z prawidłową klasyfikacją danych jawnych w całym zakresie przeanalizowanych plików, tj. od 1 B do 2323 B.



Rys. 6. Zależność różnicy odchyleń standardowych danych w funkcji ich długości dla danych krótkich
Fig. 6. Relation between difference of standard deviations and length of data for short data

4.3.3. Pierwszy moment centralny

Algorytm klasyfikacji danych jako zaszyfrowane/jawne działa w następujący sposób:

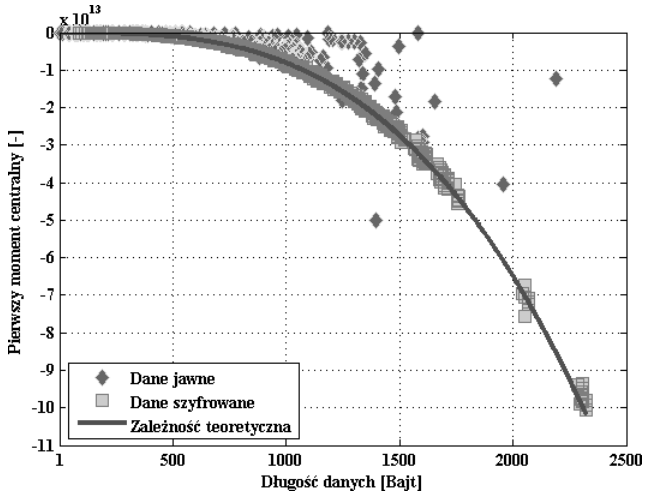
- wyznaczona zostaje wartość teoretyczna dla parametru pierwszy moment centralny przy pomocy funkcji o równaniu $f(N) = 127{,}5 \cdot (63{,}75)^m \cdot N^{2m+1}$, gdzie m to rząd momentu centralnego, a $f(N)$ to teoretyczna wartość momentu centralnego rzędu m (zależność wyznaczono w oparciu o zależność (11) dla danych losowych o rozkładzie płaskim i wartościach z przedziału od 0 do 255),
- na podstawie wszystkich bajtów danych obliczona zostaje wartość ich pierwszego momentu centralnego (według (11)),
- przypadki o wartości omawianego momentu mniejszej niż dolna granica tolerancji (przyjęta granica tolerancji wynosi 20% w stosunku do wartości teoretycznej), zakwalifikowane zostają jako dane zaszyfrowane, zaś pozostałe jako dane jawne.

Tak jak miało to miejsce poprzednio, wartość tolerancji zostaje wyznaczona w sposób doświadczalny, mający na celu maksymalizację sprawności metody.

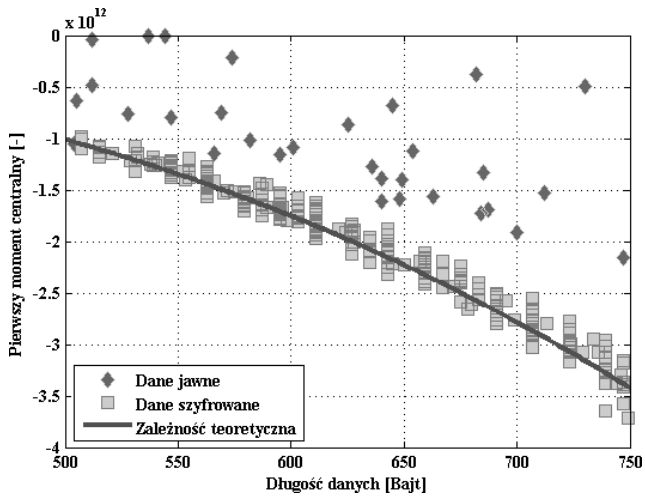
Dobrze widocznym na rysunku 7 jest skupienie wartości pierwszego momentu centralnego danych zaszyfrowanych wokół krzywej teoretycznej oraz brak tego skupienia dla danych jawnych (dla przypadków dłuższych niż 500 B).

Podobnie jak dla energii, na rysunku 8 widoczne jest również zwiększenie się odchylenia wartości momentu od zależności teoretycznej dla danych jawnych, wraz ze wzrostem długości tych danych.

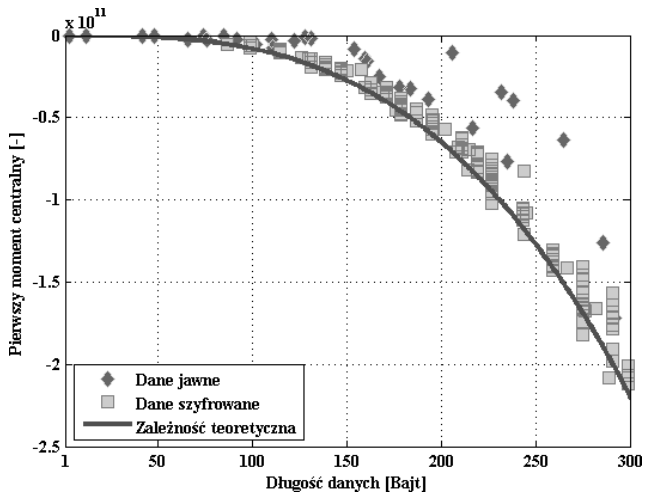
Na rysunku 9 z łatwością zauważyć można rosnące odchylenie wartości pierwszego momentu centralnego od krzywej teoretycznej dla danych jawnych, wraz ze wzrostem długości tych danych. Ponadto, podobnie jak dla przypadku z rysunku 3, zasadną wydaje się ocena, że sensowność stosowania tego parametru do określenia zaszyfrowania danych występuje dopiero w przypadku, gdy badane dane mają długość większa niż 200 B.



Rys. 7. Zależność pierwszego momentu centralnego danych w funkcji ich długości
Fig. 7. Relation between 1st central moment and length of data



Rys. 8. Zależność pierwszego momentu centralnego danych w funkcji ich długości dla wybranego przedziału długości
Fig. 8. Relation between 1st central moment and length of data for chosen data length



Rys. 9. Zależność pierwszego momentu centralnego danych w funkcji ich długości dla danych krótkich
Fig. 9. Relation between 1st central moment and length of data for short data

5. Wnioski

W artykule zaproponowano użycie wielkości charakterystycznych dla takich dziedzin wiedzy jak statystyka i teoria sygnałów, w celu detekcji informacji zaszyfrowanej. Grupa tych parametrów, zwana Parametrami Statystycznymi, charakteryzuje się efektywnością detekcji na poziomie 90% (odstaje od tego tylko jedna wielkość – wartość średnia zmodyfikowana).

Niektóre z Parametrów charakteryzują się wyjątkowo dobrą, dominującą nad innymi zdolnością do prawidłowej klasyfikacji danych jawnych (np. energia) czy też danych kodowanych (np. momenty centralne). Inne, jak odchylenie standardowe i parametry od niego pochodne, posiadają największą globalną efektywność detekcji.

Złożoność obliczeniowa zaproponowanych algorytmów, w przeważającej części przypadków, jest proporcjonalna do ich efektywności detekcji (wyjątek stanowi energia). Na przykładach od 4.3.1. do 4.3.3. pokazano, że zgodnie z tą teorią istnieje pewne ograniczenie stosowalności metod. Jest nim ilość dostępnych danych. Autorzy uważają (w oparciu o wyniki zaprezentowane powyżej oraz inne badania niezaprezentowane w niniejszej publikacji), że dolna granica zasadności stosowania metod, poniżej której liczba błędnych klasyfikacji wzrasta, a sprawność całkowita najlepszej metody spada poniżej 90%, wynosi 200 B. Granicy górnej nie stwierdzono.

Przewiduje się, że sprawność detekcji szyfrowania powinna wzrosnąć, jeśli opracowany algorytm korzystać będzie z więcej niż jednego Parametru Statystycznego. Poparciem powyższej teorii jest niecałkowite pokrycie się zbiorów poprawnie zakwalifikowanych danych dla każdego z Parametrów. Współdziałanie w ramach jednego algorytmu dwóch lub więcej Parametrów powinno doprowadzić do zwiększenia zbioru poprawnie zakwalifikowanych danych (zarówno szyfrowanych jak i nieszyfrowanych). Innymi słowy stworzenie metody łączącej w sobie Parametry posiadające najlepszą sprawność detekcji danych jawnych (np. energia) i zaszyfrowanych (np. momenty centralne), powinno doprowadzić do wzrostu efektywności detekcji szyfrowania.

Przedstawiony algorytm może być wykorzystany w aplikacjach sprzętowych i programowych służących jako analizator danych sieciowych czy też detektor przesyłu informacji szyfrowanej.

6. Literatura

[1] Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A., Dray J., Vo S.: A Statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST Special Publication 800-22, 2001.
[2] Składnikiewicz M.: Entropia – pomiar i zastosowanie, nr 3, Software Wydawnictwo, 2008. [Online]. http://www.hakin9.org
[3] Shannon C. E.: Communication Theory of Secrecy Systems, Bell System Technical Journal, vol. 28, no. 4, pp. 656-715, 1949.
[4] Arora S. i Barak B.: Cryptography, w Computational Complexity: A Modern Approach. New Jersey, Cambridge University Press, 2009, ch. 9, pp. 151-170.
[5] Barak B.: (2009, maj) Lecture 2 – Perfect Secrecy and its Limitations. [Online]. http://www.cs.princeton.edu/courses/archive/fall05/cos433/lec2.pdf
[6] Shull R.: (2004) Cryptography. [Online]. http://cs.wellesley.edu/~crypto/lectures/tr05.pdf
[7] Stinson D. R.: Cryptography : Theory and Practice, 2nd ed. Boca Raton, Chapman & Hall/CRC Press, 2002.
[8] Gajda J.: Statystyczna analiza danych pomiarowych, 1 ed. Kraków, WEAIiE AGH, 2002.
[9] Zieliński T. P.: Cyfrowe przetwarzanie sygnałów: od teorii do zastosowań, 2 ed. Warszawa, Wydawnictwa Komunikacji i Łączności, 2007.
[10] Parknet. (2009, październik) Hex-Dec Converter. [Online]. http://www.parkenet.com/apl/HexDecConverter.html

otrzymano / received: 13.05.2010
przyjęto do druku / accepted: 01.09.2010 *artykuł recenzowany*