

Efektywność parametrów statystycznych w detekcji informacji szyfrowanej

Grzegorz Gancarczyk, Agnieszka Dąbrowska – Boruch, Kazimierz Wiatr

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 Excel*.

Pytanie, na które ma dać odpowiedź niniejsze opracowanie brzmi – „czy parametry te niosą ze sobą wiedzę, którą można wykorzystać w jakiś 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.

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 niepożądanym 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.

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*”) [1]. 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 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* [2].

Stosowanie kluczy typu „*one – time pad*” jest w rzeczywistości niemożliwe ze względu na brak generatorów liczb losowych (niespełnienie warunku pełnej losowości

w pewnych przypadkach można pominąć ze względu na bardzo niską powtarzalność stosowanych algorytmów generujących ciągi liczb pseudolosowych) oraz konieczność generowania klucza o długości równej długości wiadomości jawnej [2] [3] [4] [5]. 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 [1]. „*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.** „*permutation – 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” rozumie się „o wartościach dyskretnych”).

3. PARAMETRY STATYSTYCZNE

3.1. DANE

Ponieważ w rozważaniach ujmujemy jedynie systemy szyfrujące operujące na danych binarnych w oparciu

o najpopularniejsze spośród stosowanych obecnie algorytmów szyfrujących (*3DES*, *AES*, *CAST – 128*, *RC4*, *Serpent*, *Twofish*, *Blowfish*), 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, 48 bitów, 64 bity, itd.). Poszczególne „kwanty” znajdują zastosowanie na różnych poziomach abstrakcji (bit – logika niskiego poziomu, 32 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 narzuca niejako 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 (zważywszy na fakt, że praktycznie nie spotyka się *jednostek arytmetyczno – logicznych* z magistralą mniejszą niż 8 bitów) [6] [7]. 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 wymienione wcześniej:

- 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 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, 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).

Do wykonania analiz jak i wygenerowania wyników posłużono się narzędziem numerycznym *MATLAB* firmy *Mathworks* oraz konwerterem *HEX to DEC* [8].

3.2. PARAMETRY STATYSTYCZNE

Analizę statystyczną danych przeprowadzono w oparciu o następujące parametry:

- 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 analizy danych pod kątem określenia ich zaszyfrowania 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 osiemnastu, dla których wyniki nie posiadały żadnych cech charakterystycznych, dających się zauważyć w przypadku analizy graficznej (wykresy) jak i numerycznej (wartości wyników). Cechy ten to np. skupienie wyników wokół pewnej wartości stałej (np. wariancja) lub funkcji (np. wartość średnia) czy też charakterystyczny kształt wykresu (np. histogram). Analizę przeprowadzono na ograniczonej liczbie próbek danych jawnych i zaszyfrowanych.

Tabela 1.
Parametry Statystyczne przydatne do detekcji informacji zaszyfrowanej

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

Przyglądając się parametrom z tabeli 1, dziwić może obecność takich wielkości jak *energia*, *moc średnia* czy też *moc średnia widmowa* (*wartość skuteczna* jest często spotykana w statystyce pod nazwą *wartość średniokwadratowa*). Wielkości te są raczej bardziej związane z cyfrowym przetwarzaniem sygnałów i charakterystyczne dla sygnałów niosących ze sobą informację. Jak już 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 informację.

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 zmienionej przez autorów.

3.3. DEFINICJE MATEMATYCZNE

Poniżej przedstawiono definicje matematyczne wymienionych w tabeli 1 *Parametrów Statystycznych*. Definicje niezmodyfikowane zaczerpnięto z pozycji [7].

Wartość średnia funkcji dyskretnej dana jest wyrażeniem

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

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

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

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.

Energia definiuje się jako

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

Moc średnia dana jest zależnością

$$\bar{P}_x = \frac{1}{N} \sum_{n=0}^{n=N-1} x^2(n). \quad (4)$$

Definicja matematyczna *wartości skutecznej* funkcji dyskretnej ma postać

$$RMS_x = \sqrt{\frac{1}{N} \sum_{n=0}^{n=N-1} x^2(n)}. \quad (5)$$

Wariancję danych oblicza się przy użyciu zależności

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

Możliwość wykonania modyfikacji parametru (6) wynika z dwóch sposobów interpretacji występującej w nim wartości średniej $\bar{x}$. 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

$$\sigma_x^2 = \frac{1}{N} \sum_{n=0}^{N-1} [x(n) - 127,5]^2. \quad (7)$$

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

Odchylenie standardowe dane jest zależnością jak poniżej

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} [x(n) - \bar{x}]^2}. \quad (8)$$

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).

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} [x(n) - 127,5]^2}. \quad (9)$$

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

Momenty normalne dane są jako

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

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 jakiegokolwiek próbki (sumowanie od 0 spowoduje, że pierwsza z próbek nie będzie miała żadnego wkładu do wyniku końcowego).

Momenty centralne dane są zależnością

$$\delta_x^m = \sum_{n=1}^{n=N} [n - \bar{k}_x^1]^m x(n). \quad (11)$$

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 pozostałych przypadkach zostaje obliczona wartość stosunku dla histogramu o szerokości przedziału równej 128. Jeśli stosunek dla niego znajduje się w przedziale od 0,75

do 1,25, dane zostają uznane za zaszyfrowane. Jeśli dla obu szerokości histogramu stosunki znajdują się w przedziale od 1,25 do 2,00, bada się trend zmian wartości tych stosunków. 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 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

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

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 poniżej (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.

Efektywność detekcji informacji zaszyfrowanej zostaje obliczona przy pomocy zależności

$$\eta_{\%} = \frac{100}{N} \sum_{i=1}^{i=K} \delta_i, \quad (13)$$

gdzie

$$\delta_i = \frac{c_i}{c_i + u_i}, \quad (14)$$

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, omawiane *momenty centralne* oraz *histogram zmodyfikowany*. Pozostałe, choć ich sprawność 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 [6].

Tabela 2.
Efektywność Parametrów Statystycznych w detekcji
szyfrowania informacji

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 odchyłeń standardowych	91,91
Moment normalny od zerowego do piątego	89,91
Moment centralny 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.

Z kolei *momenty centralne* 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*).

Zwrócić uwagę należy również na stosunkowo wysoką efektywność najprostszej z omawianych metod – *wartości średniej zmodyfikowanej*. Ponad 80% efektywność może nie jest wyjątkowo dobrym wynikiem, jednak z pewnością dobrym, jeśli spojrzeć na wyjątkową prostotę metody (zarówno obliczeniową jak i definicyjną).

4.3. INTERPRETACJA GRAFICZNA

W paragrafie przedstawiono w sposób graficzny rozkład wartości *energii*, *pierwszego momentu centralnego*

oraz *różnicy odchyłeń standardowych* w funkcji długości informacji testowej. W sposób wyraźny odróżniono przypadki odpowiadające informacji jawnej (symbol $\diamond$) od przykładowych próbek informacji zakodowanej (symbol $\square$). 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:

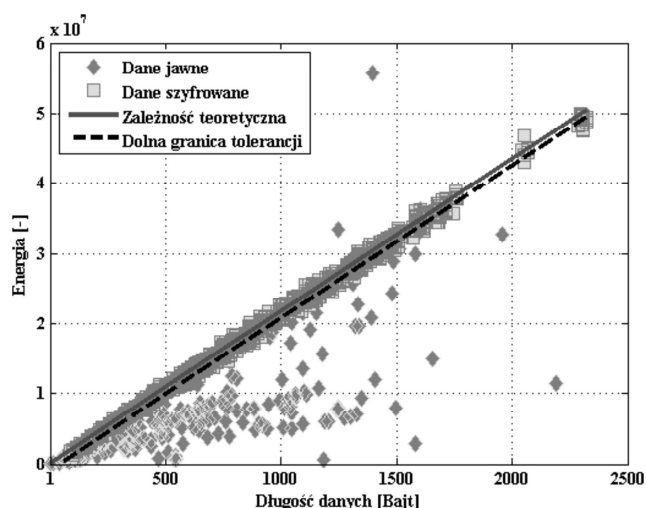
- wyznaczona zostaje zależność teoretyczna dla Parametru *energia* (funkcja o równaniu $f(N) = 21717,5 \cdot N$, gdzie N to długość danych, a $f(N)$ to teoretyczna wartość energii danych o rozkładzie płaskim i wartościach z przedziału od 0 do 255),
- wyznaczona zostaje dolna granica tolerancji dla Parametru *energia* (w omawianym przypadku jest to funkcja o równaniu $g(N) = 21717,5 \cdot (N - 42)$),
- 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 jak dokonano wyboru równania opisujące 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.

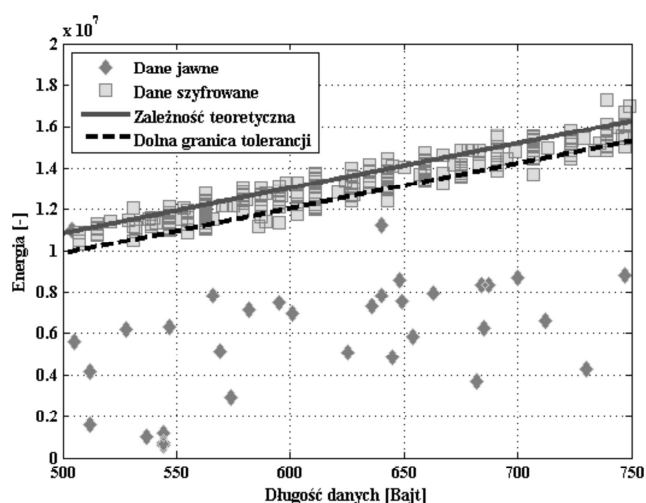
Na rysunku 1 skupienie danych wokół prostej teoretycznej jest bardzo dobrze widoczne. Kolejna zauważalna prawidłowość, to zgodne z oczekiwaniami, rosnące wraz z długością odchylenie danych jawnych od krzywej teoretycznej. Dokładnie odwrotny proces zachodzi w przypadku danych zaszyfrowanych.

Przyglądając się wykresowi z rysunku 2 zauważyć można, że błąd popełniany w przypadku danych jawnych występuje jedynie raz (okolicie 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.

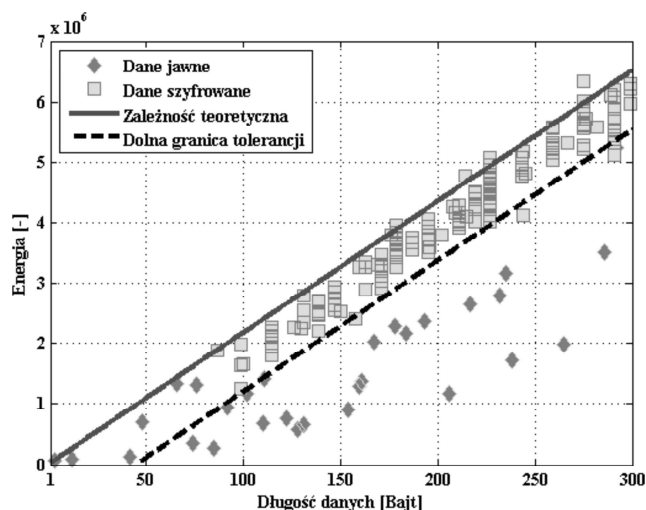
Doskonale widocznym na rysunku 3 jest, że równanie opisujące prostą będącą dolną granicą tolerancji, całkowicie uniemożliwia poprawną klasyfikację danych krótszych niż około 50 B (energia musiałaby być ujemna, aby klasyfikacja mogła być poprawna. Taki przypadek jest zgodnie z definicją (3) niemożliwy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



Rysunek 1. Zależność energii danych w funkcji ich długości



Rysunek 2. Zależność energii danych w funkcji ich długości dla wybranego przedziału długości



Rysunek 3. Zależność energii danych w funkcji ich długości dla danych krótkich

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

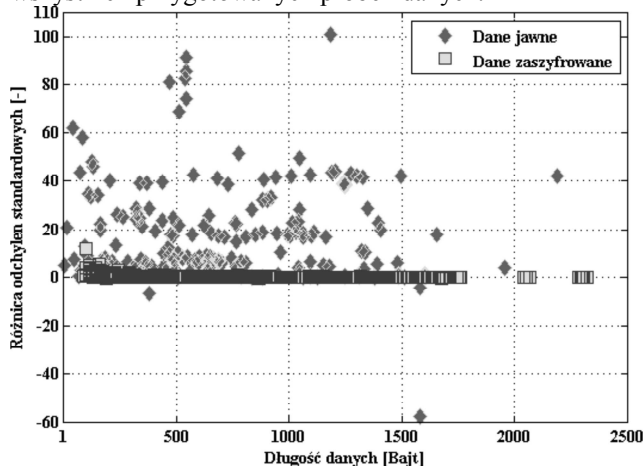
energia może zostać zastosowany w metodzie detekcji szyfrowania informacji w przypadku, gdy ma się do czynienia z informacją dłuższą niż 200 B.

4.3.2. RÓŻNICA ODCHYLEŃ STANDARDOWYCH

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

- na podstawie wszystkich bajtów danych obliczone zostaje ich *odchylenie standardowe* oraz *odchylenie standardowe zmodyfikowane* (według (8) i (9)),
- przypadki, dla których wartości *różnicy odchyłeń standardowych* nie mieszczą się wewnątrz przedziału skupionego wokół wartości 0 (przedział zawiera się w granicach od -1,20 do 2,25 wyznaczonych doświadczanie 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.



Rysunek 4. Zależność różnicy odchyłeń standardowych danych w funkcji ich długości

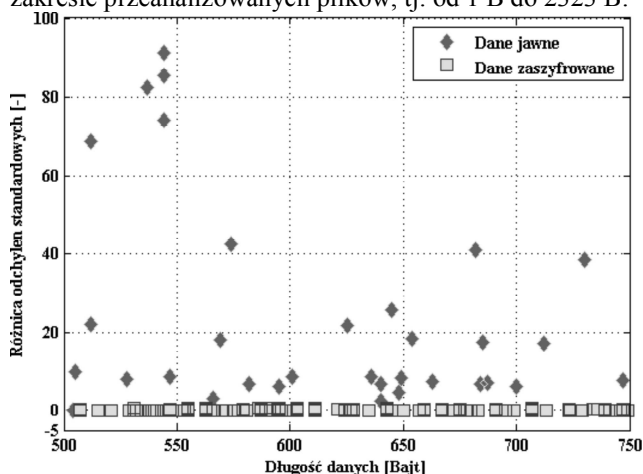
W przypadku danych zaszyfrowanych, których teoretyczna wartość średnia powinna być na poziomie 127,5, różnica odchyłeń standardowych powinna wynieść 0. Przyglądając się umiejscowieniu punktów odpowiadającym informacji tajnej zauważyć można, że trend taki występuje. Podobnie jak poprzednio, wraz ze wzrostem długości danych szyfrowanych, wzrasta ich skupienie wokół wartości teoretycznej. Dla 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.

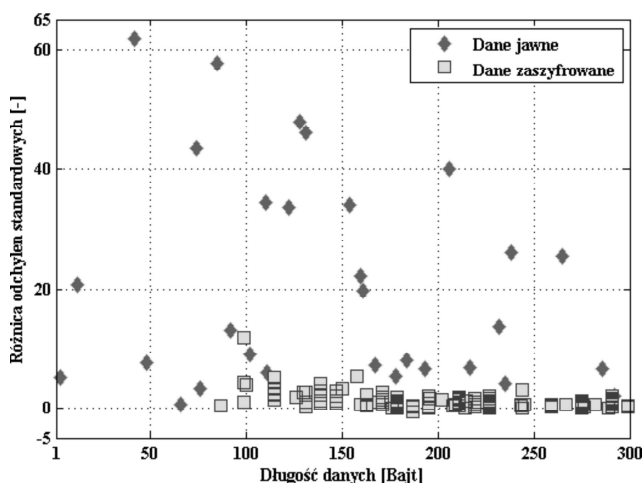
Na rysunku 5, pomimo wykonania zbliżenia na dane o określonej długości, brak jest widocznego wzrostu odchyłki od wartości teoretycznej wraz ze wzrostem długości danych. Zauważyć jednak można praktycznie stałe wartości odchyłeń dla danych zaszyfrowanych w całym analizowanym przedziale długości danych.

Widoczne na rysunku 6 znaczne odchyłki *różnicy odchyłeń standardowych* danych zaszyfrowanych sprawiają, że zasadność stosowania metody można

rozważać dopiero dla danych dłuższych niż 200 B, a nawet 250 B. Co warto zauważyć, 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.



Rysunek 5. Zależność różnicy odchylen standardowych danych w funkcji ich długości dla wybranego przedziału długości



Rysunek 6. Zależność różnicy odchylen standardowych danych w funkcji ich długości dla danych krótkich

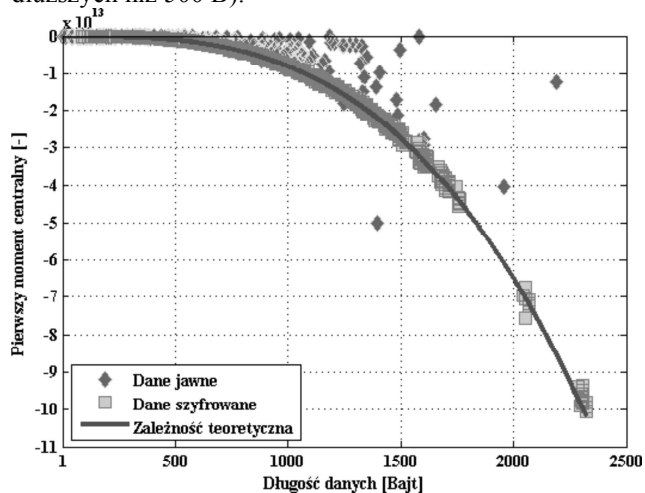
4.3.3. PIERWSZY MOMENT CENTRALNY

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

- wyznaczona zostaje zależność teoretyczna dla parametru *pierwszy moment centralny* (funkcja 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 danych o rozkładzie płaskim i wartościach z przedziału od 0 do 255),
- na podstawie wszystkich bajtów danych obliczony zostaje ich *pierwszy moment centralny* (według (11)),
- przypadki o wartości omawianego momentu mniejszej niż dolna granica tolerancji (w przypadku omawianego momentu centralnego pierwszego rzędu, granica tolerancji wynosi 20% w stosunku do wartości teoretycznej), zakwalifikowane zostają jako dane zaszyfrowane, zaś pozostałe jako dane jawne.

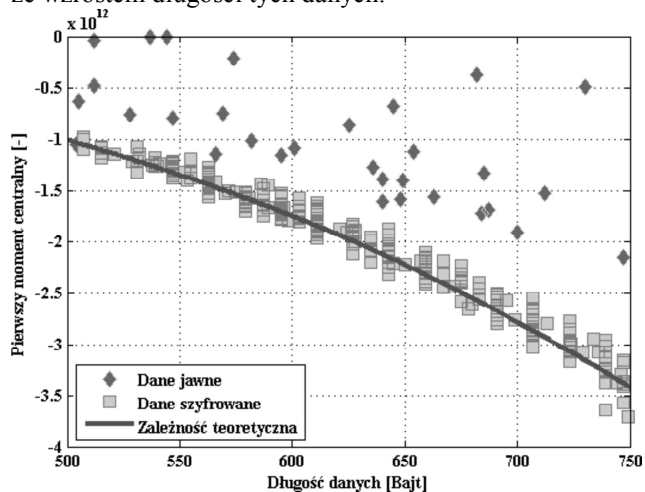
Tak jak miało miejsce poprzednio, wartość tolerancji zostaje wyznaczona w sposób doświadczalny, mający na celu maksymalizację sprawności metody.

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).



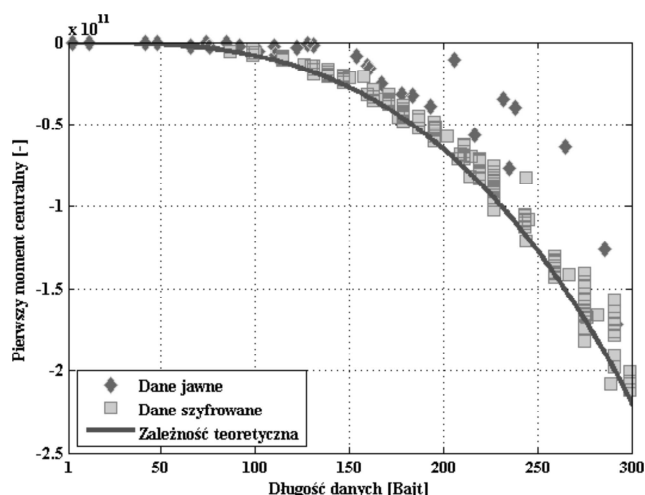
Rysunek 7. Zależność pierwszego momentu centralnego danych w funkcji ich długości

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



Rysunek 8. Zależność pierwszego momentu centralnego danych w funkcji ich długości dla wybranego przedziału długości

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 większą niż 200 B.



Rysunek 9. Zależność pierwszego momentu centralnego danych w funkcji ich długości dla danych krótkich

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 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. *Moment centralny*), 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.

LITERATURA

- [1] C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656-715, 1949.
- [2] S. Arora and B. Barak, "Cryptography," in *Computational Complexity: A Modern Approach*. New Jersey, USA: Cambridge University Press, 2009, ch. 9, pp. 151-170.
- [3] B. Barak. (2009, maj) Lecture 2 - Perfect Secrecy and its Limitations. [Online]. <http://www.cs.princeton.edu/courses/archive/fall05/cos433/lec2.pdf>
- [4] R. Shull. (2004) Cryptography. [Online]. <http://cs.wellesley.edu/~crypto/lectures/tr05.pdf>
- [5] D. R. Stinson, *Cryptography : Theory and Practice*, 2nd ed. Boca Raton, USA: Chapman & Hall/CRC Press, 2002.
- [6] J. Gajda, *Statystyczna analiza danych pomiarowych*, 1 ed. Kraków, Polska: WEAIe AGH, 2002.
- [7] T. P. Zieliński, *Cyfrowe przetwarzanie sygnałów: od teorii do zastosowań*, 2 ed. Warszawa, Polska: Wydawnictwa Komunikacji i Łączności, 2007.
- [8] Parknet. (2009, październik) Hex-Dec Converter. [Online]. <http://www.parkenet.com/apl/HexDecConverter.html>



Mgr inż. Grzegorz Gancarczyk
Akademia Górniczo – Hutnicza
Akademickie Centrum Komputerowe
Cyfronet AGH

ul. Nawojki 11
30 – 072 Kraków
e-mail:
grzegorz.gancarczyk@cyfronet.pl



Dr inż. Agnieszka Dąbrowska-Boruch
Akademia Górniczo – Hutnicza
Wydział Elektrotechniki, Automatyki,
Informatyki i Elektroniki / ACK Cyfronet
AGH
Katedra Elektroniki

ul. Mickiewicza 30 / ul. Nawojki 11
30 – 059 Kraków / 30 – 072 Kraków
e-mail: adabrow@agh.edu.pl



Prof. dr hab. inż. Kazimierz Wiatri
Akademia Górniczo – Hutnicza
Wydział Elektrotechniki, Automatyki,
Informatyki i Elektroniki / ACK
Cyfronet AGH
Katedra Elektroniki

ul. Mickiewicza 30 / ul. Nawojki 11
30 – 059 Kraków / 30 – 072 Kraków
e-mail: wiatr@agh.edu.pl