Przygotowanie i analiza danych

Wstęp

- dane z konkursu **KDD Mining Cup 1999.**

- KDD Cup jest światowym konkursem Eksploracji Danych i Odkrywania Wiedzy organizowanym przez *ACM SIGKDD - Special Interest Group on Knowledge Discovery and Data Mining.*

- **ZADANIE KONKURSU**
  - zbudowanie modelu prognostycznego (tj. klasyfikatora) zdolnego do rozróżniania połączeń “złych”, zwanych atakami lub włamaniami, oraz “dobrych” (normalnych) połączeń.

Zestaw danych

- Zestaw danych obejmuje różnego rodzaju ataki symulowane na wojskowym środowisku sieciowym

Symulowane ataki

- Denial of Service Attack (DoS - odmowa usługi)
- Remote to Local (R2L)
- User to Root (U2R)
- Probe

Liczność zbioru danych

- 24 treningowe rodzaje ataków
- 14 rodzajów ataków w danych testowych

- dane testowe nie pochodzą z tego samego rozkładu probabilistycznego jak dane testowe

- Zbiór danych zawiera:
  - 8 050 290 rekordów (1 173 MB)
  - 4 940 000 rekordów treningowych (734 MB)
  - 3 110 290 rekordów testowych (430 MB)

Opis rekordu danych

- Łącznie 42 zmienne
  - *duration* – czas połączenia w sekundach,
  - *protocol_type* – rodzaj protokołu,
  - *service* – rodzaj usługi,
  - *flag* – połączenie normalne lub błędne,
  - *src_bytes* – liczba bajtów przesłanych od źródła do celu,
  - *dst_bytes* – liczba bajtów przesłanych od celu do źródła,
  - *land* – 1 jeżeli połączenie jest z tego samego hosta lub portu, 0 w przeciwnym wypadku,
  - *wrong_fragment* – liczba błędnych fragmentów,
  - *urgens* – liczba pakietów z flagą „*urgent*”,
  - *hot* – liczba wskaźników “*hot*”,
  - *num_failed_logins* – liczba logowań zakończonych niepowodzeniem,
  - *logged_in* – *1* jeżeli użytkownik zalogowany, *0* w przeciwnym wypadku,
  - *num_compromised* – liczba skompromitowanych warunków,
  - *root_shell* – *1* jeżeli uzyskano dostęp do powłoki Root, *0* w przeciwnym wypadku,
  - *su_attempted* – *1* jeżeli użyto komendy “*su root*”, *0* w przeciwnym wypadku,
  - *num_root* liczba dostępów do konta “*Root*”,
  - *num_file_creations* – liczba operacji tworzenia pliku,
  - *num_shells* – liczba monitów powłoki,
  - *num_access_files* – liczba operacji kontroli dostępu na plikach,
  - *num_outbound_cmds* – liczba komend wychodzących w sesji *ftp*,
  - *is_host_login* – *1* jeżeli zalogowany na konto “*hot*”, *0* w przeciwnym wypadku,
  - *is_guest_login* – *1* jeżeli zalogowano na koncie gościa, *0* w przeciwnym wypadku,
  - *count* – liczba połączeń do tego samego hosta w przeciągu ostatnich 2 sekund,
  - *class* – klasa ruchu.

Opis rekordu danych

- 5 zmiennych dotyczących tego samego hosta
  - *serror_rate* – procent połączeń z błędem “*SYN*”,
  - *rerror_rate* – procent połączeń z błędem “*REJ*”,
  - *same_srv_rate* – procent połączeń do tego samego serwisu,
  - *diff_srv_rate* – procent połączeń do innego serwisu,
  - *srv_count* – liczba połączeń do tej samej usługi w przeciągu ostatnich 2 sekund.

Opis rekordu danych

- Kolejnych 13 dotyczy tej samej usługi
  - *srv_serror_rate* - procent połączeń z błędem "*SYN*",
  - *srv_rerror_rate* - procent połączeń z błędem "*REJ",*
  - *srv_diff_host_rate* – procent połączeń do innego hosta,
  - *dst_host_rate* – procent połączeń do tego samego hosta docelowego,
  - *dst_host_count* – liczba połączeń do tego samego hosta docelowego, (35) *dst_host_srv_count* – liczba połączeń do tego samego hosta docelowego i tej samej usługi,
  - *dst_host_same_srv_rate* – procent połączeń do tego samego hosta docelowego i do tej samej usługi,
  - *dst_host_diff_srv_rate* – procent różnych usług połączonych do danego hosta,
  - *dst_host_same_src_port_rate* – procent połączeń do hosta z tym samym portem źródłowym,
  - *dst_host_srv_diff_host_rate* – procent połączeń do tej samej usługi od różnych hostów,
  - *dst_host_serror_rate* – procent błędnych połączeń do hosta z błędem „*SYN*",
  - *dst_host_srv_serror_rate* – procent błędnych połączeń do hosta I tej samej usługi z błedem „*SYN*",
  - *dst_host_rerror_rate* – procent błędnych połączeń do hosta z błędem „*REJ*".

Przygotowanie danych

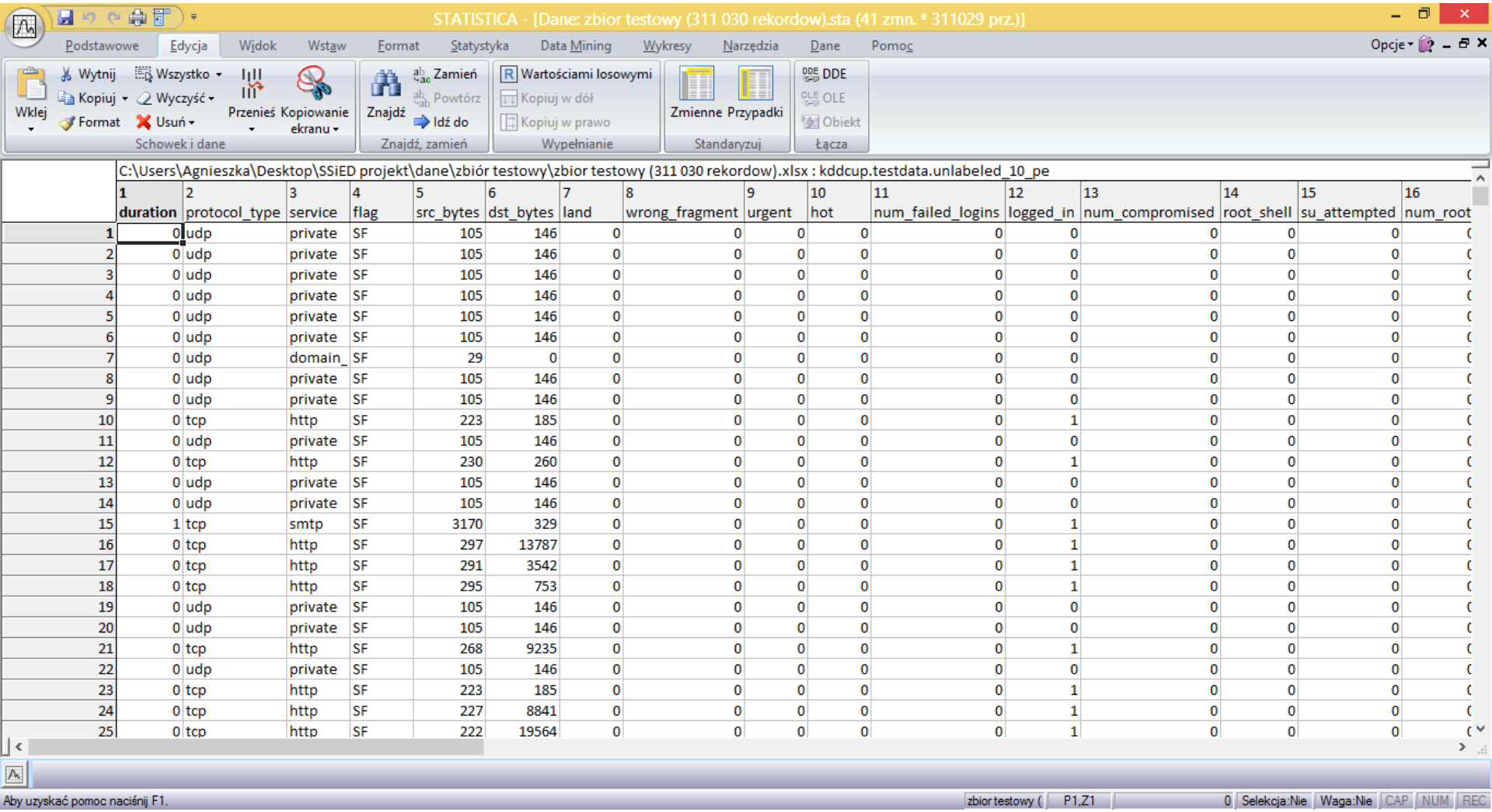
- pakiet STATISTICA 10 PL

- 10-procentowe zbiory danych ze zbiorów uczącego i testowego przygotowanego na konkurs KDD CUP 1999

  - *kddcup.data_10_percent.gz* - zbiór uczący zawierający **494 023** rekordów
  - *kddcup.testdata.unlabeled_10_percent* – zbiór testowy zawierający **311 030** rekordów

Przygotowanie danych

- odczyt danych
- przypisanie danym właściwych etykiet
- import do projektu w STATISTICA 10


STATISTICA - [Dane: zbior testowy (311 030 rekordow).sta (41 zmn. * 311029 prz.)]
Podstawowe
Edycja
Widok
Wstaw
Format
Statystyka
Data Mining
Wykresy
Narzędzia
Dane
Pomoc
Opcje
Wklej
Wytnij
Kopiuj
Format
Wszystko
Wyczyść
Usuń
Przenieś
Kopiowanie ekranu
Schowek i dane
Znajdź
Zamień
Powtórz
Idź do
Znajdź, zamień
Wartościami losowymi
Kopiuj w dół
Kopiuj w prawo
Wypełnianie
Zmienne
Przypadki
Standaryzuj
DDE
OLE
Obiekt
Łącza
C:\Users\Agnieszka\Desktop\SSiED projekt\dane\zbiór testowy\zbior testowy (311 030 rekordow).xlsx : kddcup.testdata.unlabeled_10_pe
1 duration
2 protocol_type
3 service
4 flag
5 src_bytes
6 dst_bytes
7 land
8 wrong_fragment
9 urgent
10 hot
11 num_failed_logins
12 logged_in
13 num_compromised
14 root_shell
15 su_attempted
16 num_root
1 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
2 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
3 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
4 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
5 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
6 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
7 0 udp domain_ SF 29 0 0 0 0 0 0 0 0 0 0
8 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
9 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
10 0 tcp http SF 223 185 0 0 0 0 0 1 0 0 0
11 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
12 0 tcp http SF 230 260 0 0 0 0 0 1 0 0 0
13 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
14 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
15 1 tcp smtp SF 3170 329 0 0 0 0 0 1 0 0 0
16 0 tcp http SF 297 13787 0 0 0 0 0 1 0 0 0
17 0 tcp http SF 291 3542 0 0 0 0 0 1 0 0 0
18 0 tcp http SF 295 753 0 0 0 0 0 1 0 0 0
19 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
20 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
21 0 tcp http SF 268 9235 0 0 0 0 0 1 0 0 0
22 0 udp private SF 105 146 0 0 0 0 0 0 0 0 0
23 0 tcp http SF 223 185 0 0 0 0 0 1 0 0 0
24 0 tcp http SF 227 8841 0 0 0 0 0 1 0 0 0
25 0 tcp http SF 222 19564 0 0 0 0 0 1 0 0 0
Aby uzyskać pomoc naciśnij F1.
zbior testowy (
P1,Z1
0
Selekcja:Nie
Waga:Nie
CAP
NUM
REC

Przygotowanie danych – usunięcie powtórzeń

- Usunięcie z arkusza powtarzających się rekordów

- Minimalizacja ryzyka wystąpienia problemu, w którym model nauczy się dobrze rozpoznawać jedynie obiekt najliczniej występujący w zbiorze uczącym

Przygotowanie danych – czyszczenie danych

- Usunięcie z arkusza rekordów niekompletnych

Przygotowanie danych – losowanie danych

- **losowy podzbiór przypadków** – losuje 20% podzbioru rekordów ze zbioru uczącego oraz 5% podzbioru rekordów ze zbioru testowego

Proces eksploracji danych zbioru uczącego