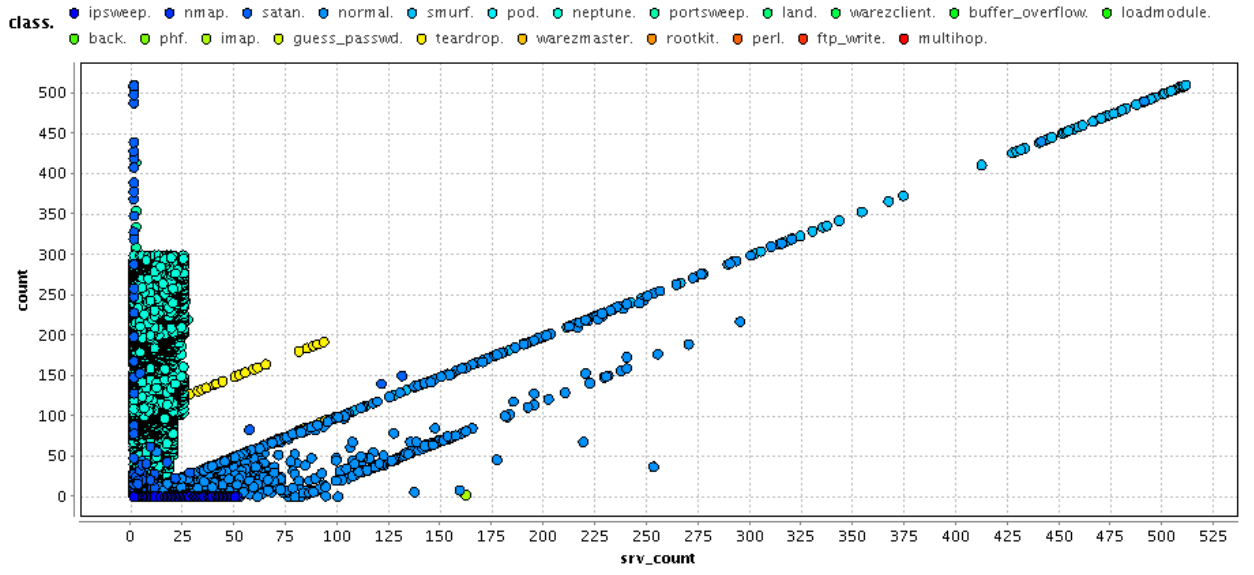
Poniżej przedstawiono wykresy zależności 2 parametrów, z wykorzystaniem zbioru uczącego zawierającego 29087 rekordów

- count - liczba połączeń do tego samego hosta w przeciągu ostatnich 2 sekund
- srv_count - liczba połączeń do tej samej usługi w przeciągu ostatnich 2 sekund

Wybór takich a nie innych parametrów spowodowany jest tym że obydwa parametry są typu liczbowego oraz że ich wartości w zbiorze danych są zróżnicowane, co umożliwia pokazania działania algorytmu SMOTE na wykresie. Dodatkowo kolor określa klasę danego pakietu.

Poniższy wykres przedstawia oryginalny, niezmodyfikowany zbiór danych:



Przynależność klas do konkretnych rodzajów ataków pokazuje poniższa lista

- probe: ipsweep, nmap, satan, portsweep, teardrop
- dos: smurf, pod, neptune, land, back
- r2l: warezclient, phf, imap, guess_passwd, warezmaster, ftp_write, multihop, spy
- u2r: buffer_overflow, loadmodule, perl, rootkit

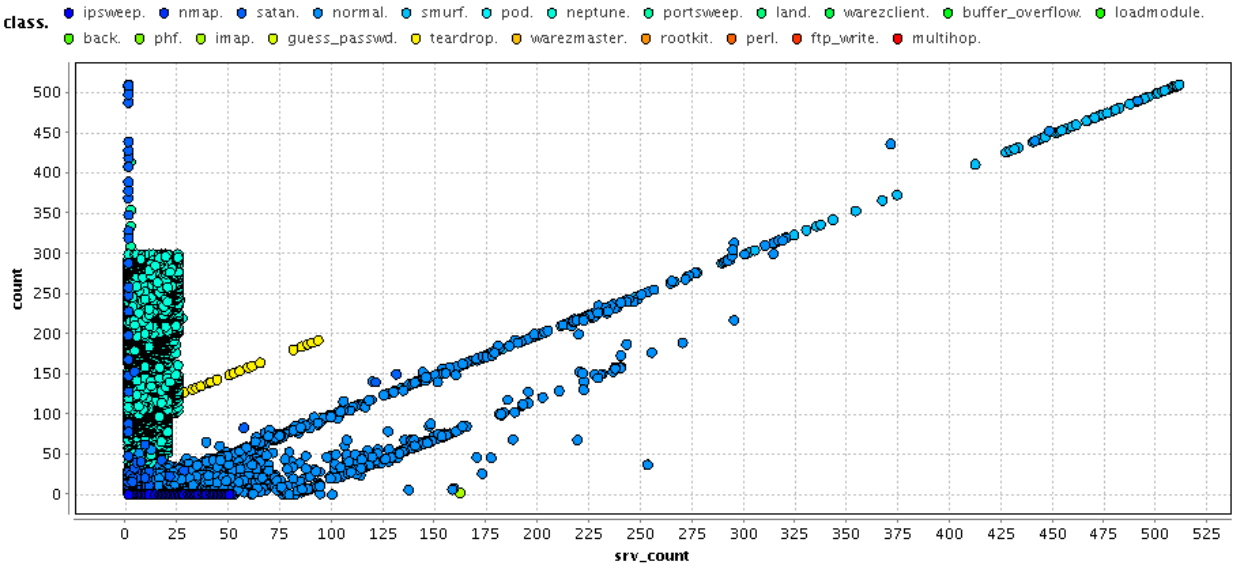
Z wykresu można stwierdzić że istnieje liczna grupa pakietów która wykazuje relacje liniową między liczbą połączeń do tego samego hosta a liczbą połączen do tej samej usługi w przeciągu ostatnich 2 sekund. Jest to grupa która składa się w większości z pakietów które nie były atakami.

Inną liczną grupe stanowią pakiety, gdzie liczba połączeń do tego samego hosta zawierała się w przedziale od 0 do 300 a liczby połączeń do usługi w przedziale od 0 do 25. Były to głównie pakiety związane z atakami typu dos.

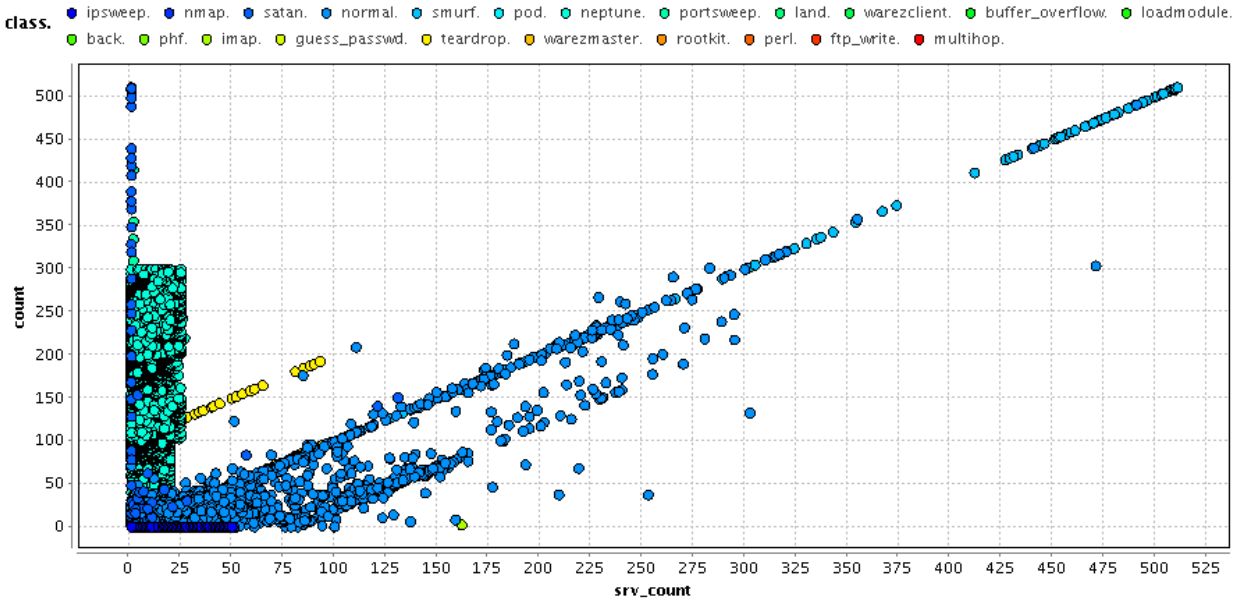
Na poniższych wykresach przedstawiono zbiór danych wzbogacony o syntetyczne dane

stworzone przy uzyciu metody SMOTE dla różnych parametrów wejściowych:
N - wartość metody SMOTE, będąca krotnością liczby 100
k - liczba najbliższych sąsiadów wziętych pod uwagę podczas tworzenia syntetycznego obiektu
Jako klase mniejszościową przyjęto klasę normal - czyli pakiety nie będące atakami.

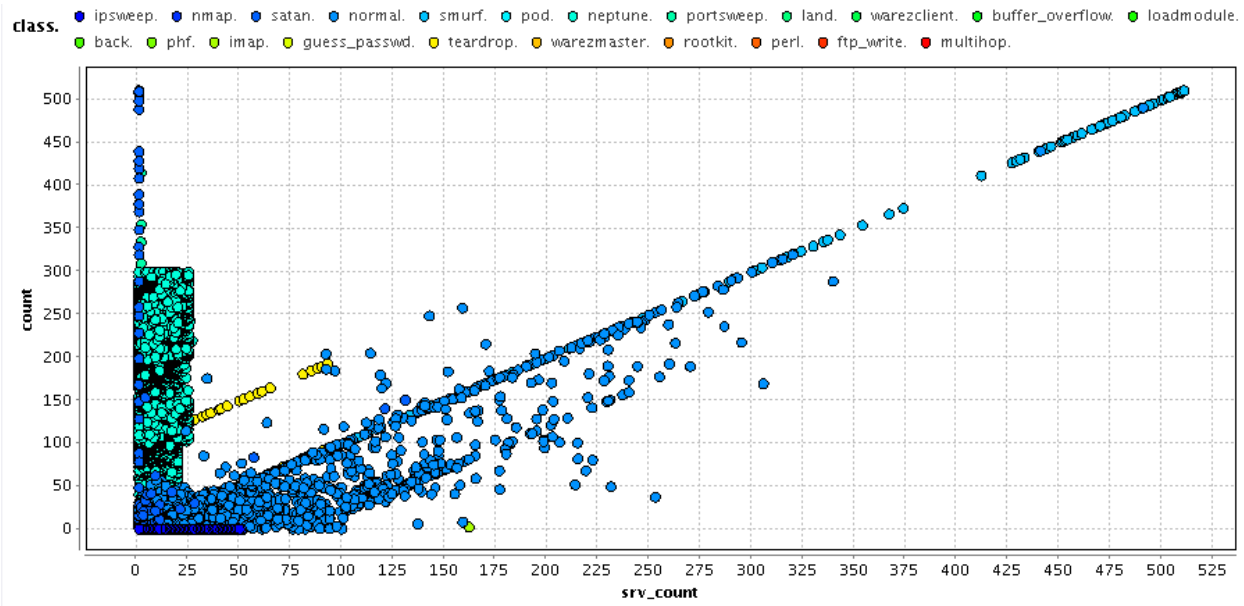
N = 100, k = 3



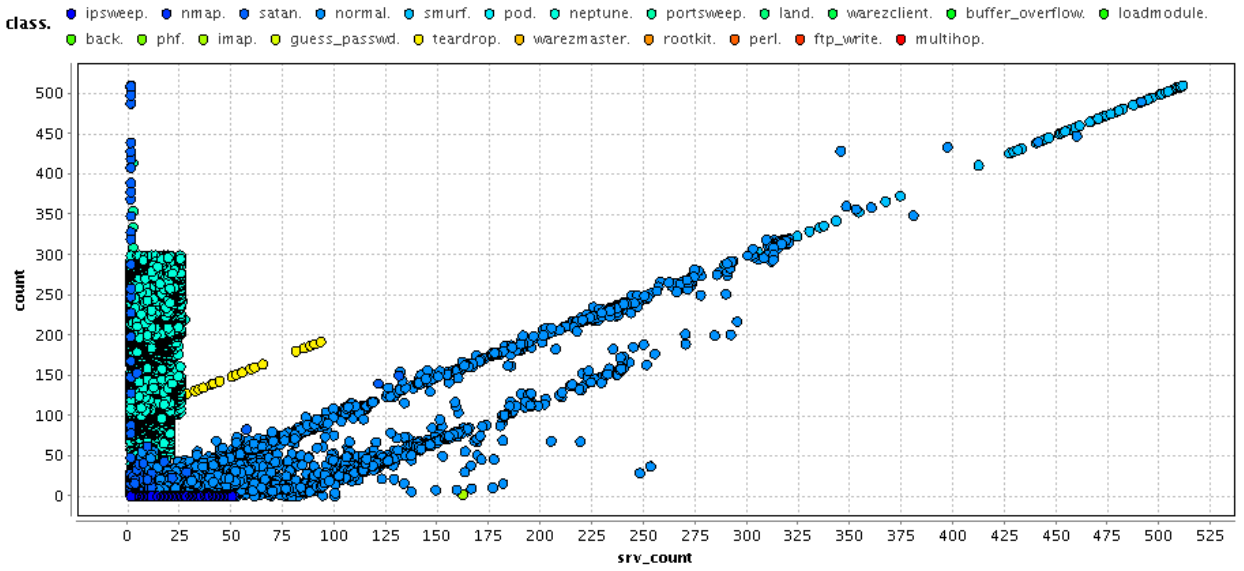
N = 100, k = 30



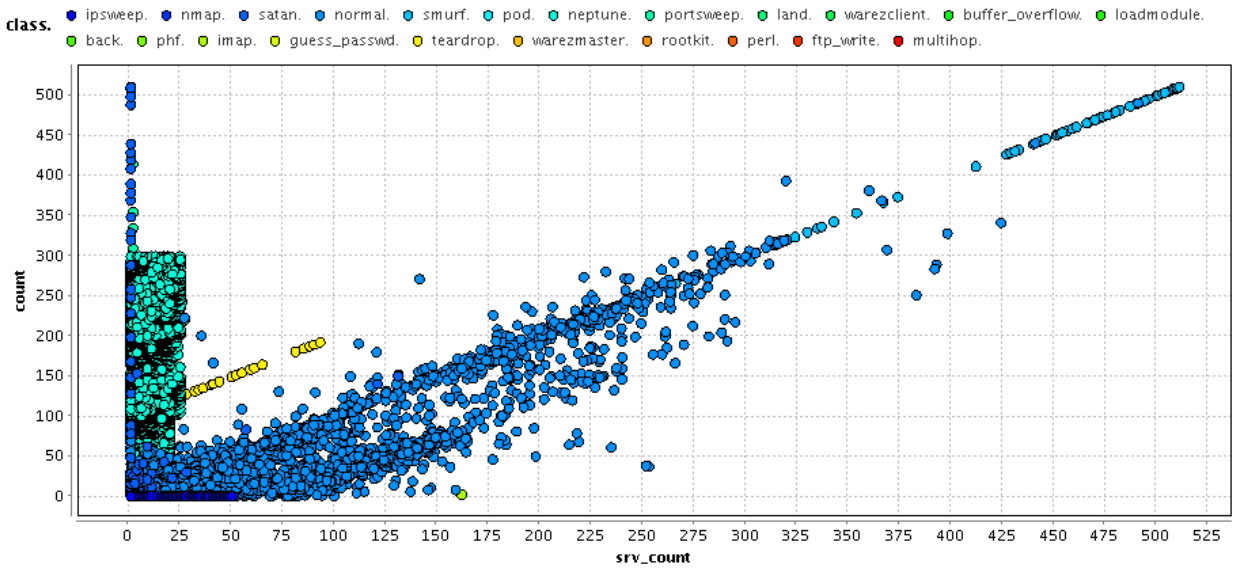
N = 100, k = 300



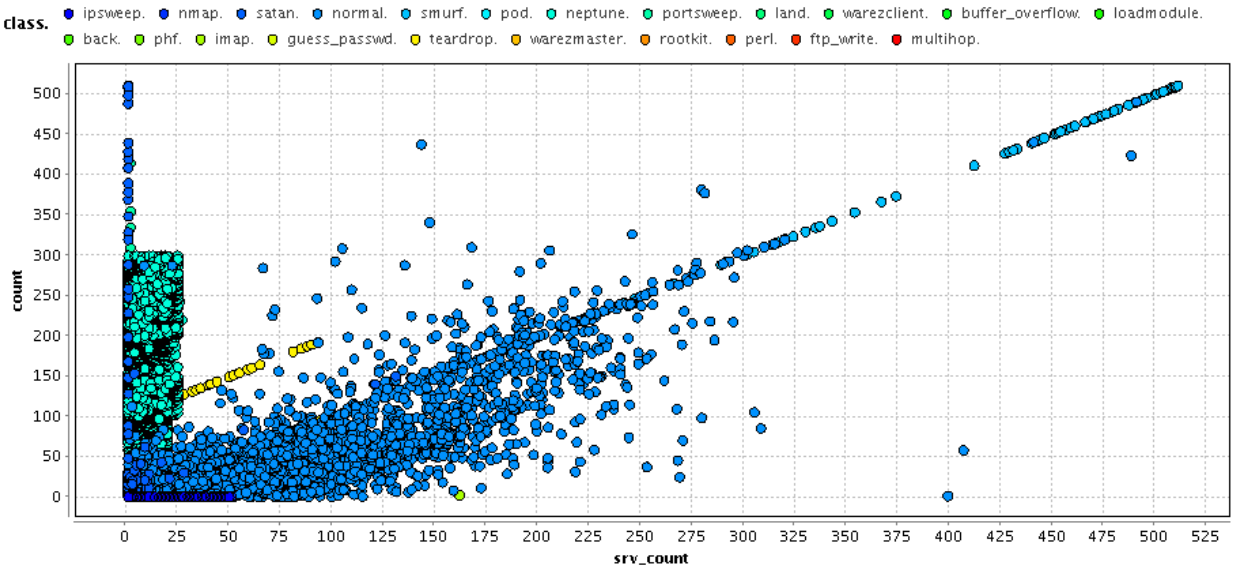
N = 500, k = 3



N = 500, k = 30



N = 500, k = 300



Na powyższych wykresach widać że algorytm działa prawidłowo, gdyż dla wyższych wartości N generuje więcej elementów, natomiast wraz ze wzrostem parametru k, elementy syntetyczne są bardziej oddalone od elementów ze zbioru danych, czyli znajdują się w dalszym sąsiedztwie.