

RACHUNEK PRAWDOPODOBIENSTWA I STATYSTYKA PROJEKT ZALICZENIOWY - RAPORT

3 maja 2016

1 Zespół

2 Pozyskanie danych

3 Opis danych

4 Czyszczenie danych

5 Agregacja i transformacja danych

Jednostka statystyczna: Pojedyncze zalogowanie.

Zbiorowość generalna: Zbiór wszystkich zalogowań do systemu.

Zbiorowość próbna: Pojedynczy użytkownik (ID).

6 Eksploracja danych

a. Liczebność względna jednostek statystycznych z niekompletnymi danymi:

$$\delta = \frac{n}{m} \quad n - \text{bldnych danych}, \quad m - \text{poprawnych danych}$$

b. Przedziały klasowe, rozpiętość przedziałów klasowych:

$$avg(x) = \frac{l_x}{d_o - d_p}$$

Gdzie: x - dany użytkownik, l_x - łączna liczba logowań użytkownika, d_o - data ostatniego logowania, d_p - data pierwszego logowania

k - ilość użytkowników

$$Kl(x) = \begin{cases} 3 & \text{dla } avg(x) > \frac{\sum_{i=0}^k avg(i)}{k} + \frac{max_{i=0..k}(avg(i)) \cdot 2}{3} \\ 2 & \text{dla } \frac{\sum_{i=0}^k avg(i)}{k} - \frac{min_{i=0..k}(avg(i)) \cdot 2}{3} \leq avg(x) \leq \frac{\sum_{i=0}^k avg(i)}{k} + \frac{max_{i=0..k}(avg(i)) \cdot 2}{3} \\ 1 & \text{dla } avg(x) \leq \frac{\sum_{i=0}^k avg(i)}{k} - \frac{max_{i=0..k}(avg(i)) \cdot 2}{3} \end{cases} \quad (1)$$

d. Wskaźnik natężenia zmian IP dla każdego użytkownika

$$W(n) = \frac{x_n}{\sum_{i=0}^k x_i} \cdot 100\%$$

e. Współczynnik korelacji pomiędzy Ilością zabezpieczeń a czasem

$$cor(X, Y) = \frac{E(XY) - EXEY}{\sqrt{D^2X \cdot D^2Y}} = \frac{\frac{\sum_{i=0}^k z_i \cdot c_i}{k} - \frac{\sum_{i=0}^k z_i}{k} \cdot \frac{\sum_{i=0}^k c_i}{k}}{\sqrt{\frac{\sum_{i=0}^k \left(z_i - \frac{\sum_{j=0}^k z_j}{k} \right)^2}{k} \cdot \frac{\sum_{i=0}^k \left(c_i - \frac{\sum_{j=0}^k c_j}{k} \right)^2}{k}}}$$

f. Wykres zmiany ilości logowań na dzień w czasie

h. Wykres z naniesionymi 3 wykresami: zmiany ilości kont danej klasy w czasie

i. Wartość minimalna funkcji prawdopodobieństwa zmiennej losowej przedstawiającej ilość logowań w danym dniu.

$$M = \min\{n_1, n_2, \dots, n_k\} \quad n_i - \text{ilosc zalogowan } i - \text{tego uzytkownika}$$

j. Prognoza obciążenia systemu dla danego dnia

$$p(mm/dd) = \frac{n(mm/dd)}{\#mm/dd}$$

- prognoza to ilość logowań w danym dniu podzielona przez ilość takich dni (jeżeli logowania miały miejsce np w różnych latach)

l. Godzina i dzień tygodnia najbardziej dogodnie do przeprowadzenia prac

$$x = \min_{dt/hh} \left(\frac{n(dt/hh)}{\#dt/hh} \right)$$

n(dt/hh) - suma logowań w danym dniu tygodnia i godzinie po wszystkich takich dniach i godzinach podzielona przez ich ilość

m. Rozkład Poissona ilości połączeń na godzinę – prawdopodobieństwo, że w danym czasie będzie dokładnie n równoczesnych połączeń:

$$P(X = k) = \frac{\frac{n(dt/hh)}{\#dt/hh}^k \cdot e^{-\frac{n(dt/hh)}{\#dt/hh}}}{k!}$$

m. Sprawdzenie czy konto x wygasł: 1 - wygasł, 0 - nie:
 d_a - aktualna data, d_o - data ostatniego

$$W(x) = \begin{cases} 1 & dla \quad d_a - d_o < 182 \\ 0 & wpp \end{cases} \quad (2)$$

7 Hipotezy badawcze

8 Weryfikacja hipotez

9 Wnioski

10 Uwagi