Informatyka stosowana, studia dzienne, II st. semestr II

Analiza danych złożonych **2020/2021**

Prowadzący: dr hab. inż. Agnieszka Duraj poniedziałek, 08:30

Referat wykładowy – Algorytm detekcji dryftu KSWIN

Autorzy:

Paweł Galewicz 234053

Karol Podlewski 234106

KSWIN (Kolmogorov-Smirnov WINdowing) to algorytm służący do detekcji dryftu oparty na statystycznym teście Kołmogorowa-Smirnowa.

Test Kołmogorowa-Smirnowa jest testem nieparametrycznym, służącym do weryfikacji hipotezy o zgodności dwóch rozkładów empirycznych ze sobą. Może być stosowany między innymi do sprawdzenia zgodności rozkładu analizowanej zmiennej z rozkładem teoretycznym (na przykład normalnym, dwumianowym). Dobrze sprawdza się przy większych próbach badawczych - częstym proponowanym przedziałem jest N $>$ 100. Istnieje też wersja testu służąca do porównania rozkładów dwóch zmiennych losowych [1].

Algorytm KSWIN, zaproponowany przez Raaba, Heusingera i Schleifa w [2], opiera się właśnie na teście Kołmogorowa-Smirnowa dla dwóch zmiennych. Potrzebuje on przesuwalnego okna Ψ o stałym rozmiarze n, z którego tworzy dwa mniejsze okna na potrzeby przeprowadzenia testu. Pierwsze z nich posiada r najnowszych punktów z okna Ψ:

$$R = \{x_i \, \epsilon \, \Psi\}_{i=n-r+1}^{n} \tag{1}$$

Drugie okno jest tworzone poprzez jednolite próbkowanie z pozostałej części okna $n - r$:

$$W = \{x_i \, \epsilon \, \Psi | i < n - r + 1 \wedge p(x) = U(x_i|1, n - r)\} \tag{2}$$

gdzie $U(x_i|1, n - r) = \frac{1}{n-r}$ jest rozkładem jednostajnym.

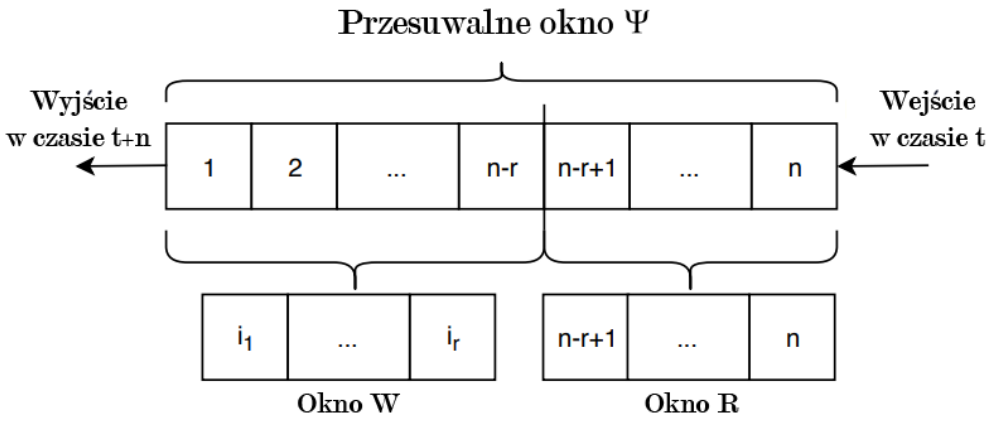
Mając już dwie zmienne, możliwe jest przeprowadzenie testu Kołmogorowa-Smirnowa, który porównuje dystrybuanty empiryczne F_W oraz F_R:

$$dist_{w,r} = \sup_x |F_W(x) - F_R(x)| \tag{3}$$

gdzie $F_{(.)}(x) = \frac{1}{n} \sum_{j=1}^{n} I_{[-\infty,x]}(X_j)$ oraz $I_{[-\infty,x]}(X_j)$ to funkcja charakterystyczna, która przyjmuje wartość 1 jeżeli $X_i \geqslant x$ lub 0 w przeciwnym wypadku. $sup(x)$ to najmniejsza wartość, dla której warunek jest prawidłowy.

Jeżeli kres górny jest większy od statystyki testowej, to hipoteza zerowa jest odrzucona, z poziomem istotności α. Dryft jest znaleziony, kiedy hipoteza zerowa zostaje odrzucona - punkt wykrycia dryftu ustawiany jest na $n - r$. Następnie okno redukowane jest do ostatnich r próbek.

Ze względu na czułość parametru α nie powinien być on większy niż 0,01 - będzie to skutkować zbyt częstym wykrywaniem dryftu.



Rysunek 1. Graficzna demonstracja podziału przesuwalnego okna Ψ

Dla okien W i R, które mają identyczny rozmiar, można zredukować test do postaci

$$dist_{w,r} > c(\alpha)\sqrt{\frac{w+r}{wr}} = \sqrt{-\frac{1}{2}\ln\alpha}\sqrt{\frac{w+r}{wr}} \overset{(w=r)}{=} \sqrt{-\frac{\ln\alpha}{r}}, \tag{4}$$

gdzie α to poziom istotności testu, r oraz w to rozmiary okien R oraz W.$c(\alpha)$ to krytyczna wartość testu, która może zostać dostosowana do standardowych rozmiarów testu lub przyjęta jako $\sqrt{-\frac{1}{2}\ln\alpha}$.

Literatura

[1] *SPSS Kolmogorov-Smirnov Test for Normality,* Ruben Gert van den Berg. https://www.spss-tutorials.com/spss-kolmogorov-smirnov-test-for-normality/ [dostęp: 04.01.2021]

[2] *Reactive Soft Prototype Computing for Concept Drift Streams,* Christoph Raab, Moritz Heusinger, Frank-Michael Schleif. https://www.sciencedirect.com/science/article/abs/pii/S0925231220305063?via%3Dihub [dostęp: 04.01.2021]