Urząd Statystyczny w Poznaniu

Klasyfikacja produktów do COICOP z wykorzystaniem machine learning

Maciej Beręsewicz, Jakub Ratajczyk, Tomasz Klimanek

Urząd Statystyczny w Poznaniu

18-19 marca 2019, Łódź

Plan prezentacji

1. Wstęp
2. Zaproponowane rozwiązanie
3. Wstępne wyniki
4. Podsumowanie
5. Literatura

Wstęp

- Celem projektu jest wypracowanie metod przypisania produktów z trzech sieci handlowych do klasyfikacji COICOP.
- W ramach klasyfikacji zakres produktów został zawężony do 6 podklas produktów żywnościowych, są to: **cukier**, **jogurt**, **kawa**, **mąka**, **mleko** oraz **ryż**.

Wstęp

- Sposoby klasyfikacji:
  - **EAN** (European Article Number), jest to rodzina kodów kreskowych wprowadzona przez stowarzyszenie European Article Numbering.
  - **COICOP**, opisuje klasyfikację spożycia według celu, pozwala na bardziej dokładnie ważenie w obliczeniach krajowych wskaźników cen towarów i usług konsumpcyjnych (CPI) jak i zharmonizowanych (HICP).
- **Indeks cen konsumpcyjnych**, stanowi relację cen reprezentatywnego zestawu dóbr konsumpcyjnych w kolejnych latach badania do ceny tego koszyka dóbr w roku bazowym, czyli przyjętym za podstawę wyliczeń.

Wstęp

- COICOP stosowany jest w wielu różnych obszarach statystyki, np. rachunek narodowy.
- Obecnie stosowany stopień klasyfikacji składa się z pięciu szczebli:
  - dwucyfrowego – 12 działów,
  - trzycyfrowego – 44 grup,
  - czterocyfrowego – 110 klas,
  - pięciocyfrowego – 296 podklas.

Dane

- Dane dostarczyły 3 sieci handlowe.
- Każda sieć posiada własny sposób opisu oraz klasyfikacji produktów, więc dla każdej sieci wskazane było użycie innego podejścia.
- Sieć 1 dostarczyła dane zawierające następujące informacje: kod hierarchii, nazwa dla kodu hierarchii, kod artykułu oraz nazwa artykułu.
- Sieć 2 dostarczyła następujące informacje: kod hierarchii oraz nazwę produktu oraz opis kodów hierarchii.
- Sieć 3 dostarczyła informacje na temat kodu hierarchii, kodu produktu oraz jego nazwy.

Dane

<chr>	<chr>	<chr>	<chr>	<chr>
40103	jogurty pitne	13938	del/2/B/DANAlejogurtowy truskawka700gbut	5900643028896
40103	jogurty pitne	13938	del/2/B/DANAlejogurtowy truskawka700gbut	5900643028919
40103	jogurty pitne	13938	del/2/B/DANAlejogurtowy truskawka700gbut	5900643032442
40103	jogurty pitne	13938	del/2/B/DANAlejogurtowy truskawka700gbut	5900643032459
40103	jogurty pitne	25383	G/Activia do picia malina/granat 300g	5900643029831
40103	jogurty pitne	25383	G/Activia do picia malina/granat 300g	5900643029855

Rys. 1: Przykładowe dane z sieci 2

V3 <chr>	V4 <chr>	V5 <chr>
104010101	236971	_BUT. PIWO ŻUBR 0,5L
121010101	210250	WORECZKI ROLOW. Z BOBINĄ /ROLKA X200 SZT
141010106	285895	ZFŚS WYPRAWKA SZKOLNA 2016
211010103	156159	CUKIER SŁODKA ŁYŻECZKA 1KG BIEDRONKA
211010104	236179	CUKIER TRZCINOWY 500G
211010104	241585	CUKIER TRZCINOWY DIAMANT 0,5KG
211010104	236179	CUKIER TRZCINOWY KRÓLEWSKI 500G

Rys. 2: Przykładowe dane z sieci 3

Zidentyfikowane problemy

- Brak zbioru uczącego.
- Opis produktów zawiera wiele informacji, które są zbędne lub niepełne (np. skróty).
- Opis produktów czasem jest niejednoznaczny.
- Opis produktów zawiera również nazwy własne (np. marki).

Zaproponowane rozwiązanie

Zaproponowane podejście oparte było na następujących krokach

1. Manualnym mapowaniu klasyfikacji sieci handlowych do COICOP (tam gdzie była taka możliwość).
2. Mapowanie służyło do utworzenia zbioru *uczącego i testowego*.
3. Czyszczeniu i ujednoliceniu nazw produktów między sieciami handlowymi.
4. Zastosowaniu algorytmów uczenia maszynowego do klasyfikacji produktów na podstawie nazw i charakterystyk do odpowiednich kategorii COICOP.

Przetwarzanie tekstu

Do przetwarzania tekstu wykorzystano

1. **wyrażenia regularne**, które umożliwiły wyczyszczenie nazw produktów. Przykładowo, poniższy kod wyszukiwał informacji o gramaturze, pojemności czy liczbie sztuk

```
(\\d{1,}(\\,|\\.|X|x|\\/))?\\d{1,}(\\s|\\.|x)?
(G|GR|gr|g|ML|ml|L|l|kg|KG|Kg|SZT|szt|MG|mg|
TB|tb|but|But|BUT|gbut|gGut)+
```

2. **stemming**, który polega na wydobyciu z wybranego wyrazu tzw. rdzenia,
3. **lematyzację** czyli sprowadzenie grupy wyrazów stanowiących odmianę danego zwrotu do wspólnej postaci, umożliwiającej traktowanie ich wszystkich jako te samo słowo.

Uczenie maszynowe – idea

- Nie zawsze dysponujemy hierarchią produktów sieci handlowych czy Europejskim Kodem Towarowym (ang. ang. European Article Number; EAN).
- W takim przypadku produkty należy przypisać do określonej grupy COICOP na podstawie ich nazwy, jak na poniższym przykładzie

Produkt	**Grupa COICOP**
jogurt truskawka 700g	?
kawa zbożowa inka	?
mleko UHT	?

- Produkty mogą być zapisane w różny sposób oraz zawierać wiele niepotrzebnych informacji.
- Ta procedura jest kluczowa jeżeli celem będzie również wykorzystywanie web-scrappingu.

Uczenie maszynowe

- W ramach projektu postanowiono zweryfikować użyteczność uczenia maszynowego do mapowania produktów na podstawie nazw do klasyfikacji COICOP.
- W tym celu wykorzystano algorytm LASSO przy założeniu rozkładu wielomianowego ponieważ klasyfikowano produkty do 9 kategorii.

Regresja LASSO – idea

- Regresja LASSO została zaproponowana przez Santosa i William (1986) oraz Tibshirani (1996). Jest szczególnym przypadkiem regresji Elastic-Net (EN; Zou i Hastie, 2005).
- Pakiety statystyczne (m.in. glmnet) minimalizują następującą funkcję wiarygodności dla regresji EN

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right], \tag{1}$$

- gdzie $l()$ to funkcja ujemnej log-wiarygodności, β to parametry modelu, x_i to zmienne, a w to wagi dla obserwacji.
- Jeżeli $\alpha = 1$ otrzymujemy regresję LASSO, a jeżeli $\alpha = 0$ otrzymujemy regresję Ridge.

Regresja LASSO – model wielomianowy

- W przypadku modelu wielomianowego, załóżmy, że badana zmienna ma K poziomów z $\mathcal{G} = \{1, 2, \ldots, K\}$ wtedy model ma postać

$$\Pr(G = k|X = x) = \frac{e^{\beta_{0k}+\beta_k^T x}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_\ell^T x}}. \tag{2}$$

Regresja LASSO – model wielomianowy

- Niech Y będzie macierzą indykatorową $N \times K$ z elementami $y_{i\ell} = I(g_i = \ell)$. Wtedy, funkcja wiarygodności dla regresji elastic-net ma postać

$$\ell(\{\beta_{0k}, \beta_k\}_1^K) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{k=1}^{K} e^{\beta_{0k}+x_i^T\beta_k}\right)\right)\right] + \lambda\left[(1-\alpha)||\beta||_F^2/2 + \alpha\sum_{j=1}^{p}||\beta_j||_q\right]. \tag{3}$$

- Gdzie β jest macierzą współczynników o wymiarach $p \times K$, β_k odnosi się do k-tego poziomu Y, a β_j odnosi się do j-tego wiersza (wektora K parametrów zmiennej j)

Regresja LASSO – model wielomianowy

- Dobór modelu oparty był podstawie wyników 10-krotnej walidacji krzyżowej (ang. *10-fold cross-validation*).
- Miarą straty wykorzystywaną do określenia parametrów był błąd błędnej klasyfikacji (ang. *misclassification error*).
- W obliczeniach wykorzystano R oraz pakiet `glmnet`.

Mapowanie kodów COICOP

Tab. 1: Wynik mapowania kodów COICOP

Sieć handlowa	Czy przypisano kod COICOP?	Liczba rekordów
Sieć 3	NIE	7499
	TAK	3311
Sieć 2	TAK	2174
Sieć 1	TAK	877

Uczenie maszynowe

Zbiór danych wykorzystany do uczenia maszynowego (na jego podstawie utworzono zbiór uczący oraz testowy w trakcie walidacji krzyżowej).

Sieć handlowa	Liczba produktów
Sieć 3	341
Sieć 2	838
Sieć 1	665
Razem	1844

Uczenie maszynowe

Wyniki klasyfikacji z wykorzystaniem regresji LASSO (kolumny – wartość prawdziwa, wiersze – wynik klasyfikacji)

	Ryż	Mąka	Mąki.P	M.Pełne	M.Nisko	M.Zagę	Jog	Cuk	Kawa
Ryż	147								
Mąka		103							
Mąki.Poz		1	48						
M.Pełne				57	3				
M.Nisko					87		1		
M.Zagę						38			
Jogurty							705		
Cukier								68	
Kawa									586

Podsumowanie

Do najważniejszych wniosków należy wymienić

- Wykorzystanie nazw produktów pozwala na bardzo dobre klasyfikowanie do COICOP.
- Znajomość słownika używanego przez sieci handlowe jest niezbędne do utworzenia zbioru uczącego oraz mapowania COICOP.
- Kluczowe jest ujednolicenie nazw produktów uwzględniając wszystkie sieci handlowe jednocześnie.
- Dalsze prace dotyczą m.in. uwzględnienia tych słów lub kategorii, których algorytm nie widział.

Wybrana literatura

- Santosa, Fadil; Symes, William W. (1986). **Linear inversion of band-limited reflection seismograms**. SIAM Journal on Scientific and Statistical Computing. SIAM. 7 (4): 1307–1330. doi: 10.1137/0907
- Friedman, Jerome; Hastie, Trevor; Tibshirani, Robert. 2010. **Regularization Paths for Generalized Linear Models via Coordinate Descent**. Journal of Statistical Software 33 (1): 1-21.
- Tibshirani, Robert (1996). **Regression Shrinkage and Selection via the lasso**. Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88.
- Zou, Hui; Hastie, Trevor (2005). **Regularization and Variable Selection via the Elastic Net**. Journal of the Royal Statistical Society. Series B (statistical Methodology). Wiley. 67 (2): 301–20. doi:1 0.1111/j.1467-9868.2005.00503.x.

Urząd Statystyczny w Poznaniu

Dziękuję za uwagę