Zastosowanie nowoczesnych technologii informatycznych w korzystaniu z alternatywnych źródeł danych

Klasyfikacja i dopasowanie produktów

Program NCBiR GOSPOSTRATEG

Projekt INSTATCENY

Zespół Podstaw Sztucznej Inteligencji IPI PAN

Warszawa, 14 czerwca 2022

Plan prezentacji

Klasyfikacja produktów

Klasyfikator produktów

- Trenowanie klasyfikatora produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu (można też uwzględnić dodatkowe cechy)
- Zarządzanie modelami wytrenowanych klasyfikatorów przechowywanymi na dysku
- Klasyfikacja produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu lub z uwzględnieniem dodatkowych cech.

Dane

Format danych treningowych / do klasyfikacji

- Wejście stanowi nazwa pliku wraz z jego ścieżką w formacie csv.
- Plik musi zawierać pole *desc* zawierające opis produktu.
- Ponadto musi zawierać inne pola, jeśli mają być użyte w procesie trenowania.

	time	prices	quantities	codeIN	codeOUT	desc
0	2020-12-31	7.18	35.7994428969359	204189	5906340630196	cukier biały kostka n 500g
1	2020-12-31	7.38	39.3794037940379	204189	5906340630196	cukier biały kostka n 500g
2	2020-12-31	7.18	62.650417827298	204189	5906340630196	cukier biały kostka n 500g
3	2020-12-31	10.98	0.895264116575592	519581	3596710466771	>>cukier biały w sasz.100x5g
4	2020-12-31	10.98	4.47540983606557	519581	3596710466771	>>cukier biały w sasz.100x5g
5	2020-12-31	10.98	1.7896174863388	519581	3596710466771	>>cukier biały w sasz.100x5g
6	2020-12-31	7.18	48.3300835654596	204189	5906340630196	cukier biały kostka n 500g
7	2020-12-31	7.38	7.15989159891599	204189	5906340630196	cukier biały kostka n 500g
8	2020-12-31	7.98	26.8496240601504	204189	5906340630196	cukier biały kostka n 500g
9	2021-03-31	4.36	79.75	765753	5904215140252	>>KEFIR W BUTELCE 250G
10	2021-03-31	4.36	79.75	765753	5904215140252	

Algorytmy

Metody uczenia maszynowego dostępne do klasyfikacji:

- Regresja logistyczna

  `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

- Naiwny Bayes

  `https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB`

- Lasy Losowe

  `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

- liniowa maszyna wektorów nośnych (LinearSVC)

  `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

Przykład zastosowania wytrenowanego modelu do klasyfikacji

Wynikowa ramka danych

	time	prices	quantities	codeIN	codeOUT	desc	coicop_predicted
0	2020-12-31	7.18	35.7994428969359	204189	5906340630196	cukier biały kostka n 500g	c011811_1
1	2020-12-31	7.38	39.3794037940379	204189	5906340630196	cukier biały kostka n 500g	c011811_1
2	2020-12-31	7.18	62.650417827298	204189	5906340630196	cukier biały kostka n 500g	c011811_1
3	2020-12-31	10.98	0.895264116575592	519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
4	2020-12-31	10.98	4.47540983606557	519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
5	2020-12-31	10.98	1.7896174863388	519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
6	2020-12-31	7.18	48.3300835654596	204189	5906340630196	cukier biały kostka n 500g	c011811_1
7	2020-12-31	7.38	7.15989159891599	204189	5906340630196	cukier biały kostka n 500g	c011811_1
8	2020-12-31	7.98	26.8496240601504	204189	5906340630196	cukier biały kostka n 500g	c011811_1
9	2021-03-31	4.36	79.75	765753	5904215140252	>>KEFIR W BUTELCE 250G	c011462_2
10	2021-03-31	4.36	79.75	765753	5904215140252		c121321_22

Dopasowywanie (matching) produktów

Procedura dopasowywania (matching) produktów:

- Na wejściu pobierana jest ramka danych z następującymi kolumnami: *['seller_code', 'ean', 'desc']* – czyli *kod sprzedawcy*, *ean* i *opis produktu*.
- Na wyjściu zwracany jest ramka danych z dodatkową kolumną o nazwie **group**. W kolumnie tej dla każdej obserwacji znajduje się numer, który określa przynależność do grup stworzonych w wyniku matchingu. Grupy te numerowane są kolejnymi liczbami naturalnymi $(0, 1, 2, \dots)$.
- Stworzono dwie implementacje powyższej procedury.
- Oba algorytmy bazują na koncepcji porównania par zadanych produktów.

Dopasowywanie (matching) produktów – dane wejściowe

seller_code	ean	desc	coicop
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
204189	5906340630196	cukier biały kostka n 500g	c011811_1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1

Dopasowywanie produktów bazuje na poniższej procedurze porównywania produktów parami:

1. Za dopasowane produkty uznajemy te, które mają ten sam kod kreskowy (EAN) i kod produktu według sprzedawcy.
2. W przypadku produktów, które mają ten sam kod produktu według sprzedawcy, ale różny kod EAN, za dopasowane uznajemy te produkty, które mają tę samą etykietę produktu lub etykietę bardzo podobną. Za podobne uważane są produkty, dla których podobieństwo pomiędzy ich opisami jest większe od zadanego parametru.
3. W przypadku, gdy produkty mają różny kod sprzedawcy i różne kody EAN, za dopasowane uznamy te produkty, które mają identyczne etykiety.

Dopasowywanie produktów: wersja o liniowej złożoności

Metoda nie wylicza podobieństwa produktów metodą „każdy z każdym", przez co działa szybko (liniowa złożoność względem liczby danych)

Algorytm przebiega następująco:

1. produkty przetwarzamy po kolei
2. dany produkt porównujemy po kolei z już stworzonymi grupami (UWAGA: porównanie z grupą polega na porównaniu z tylko jednym jej elementem, który nazywamy jej „reprezentantem")
3. jeżeli produkt nie jest podobny do żadnej z już istniejących grup, wtedy tworzona jest dla niego nowa grupa – zaś produkt ten staje się jej reprezentantem.

Dopasowywanie produktów: wersja dokładna (kwadratowa złożoność)

Metoda ta posiada kwadratową złożoność (działa wolno dla dużych danych), jednakże jest dokładniejsza

Algorytm przebiega następująco:

1. Podobieństwa dla produktów wyliczane są metodą „każdy z każdym".
2. Następnie wybiera się te grupy produktów, które tworzą spójne składowe.
3. Dla przyspieszenia całej procedury obliczanie podobieństw wykonywane jest w sposób równoległy na wielu rdzeniach procesora.

Przykład zastosowania algorytmu do dopasowywania produktów

Wynikowa ramka danych

seller_code	ean	desc	coicop	group
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1	1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1	1
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1	1
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1	1
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0
519581	3596710466771	>>cukier biały w sasz.100x5g	c011811_1	1
204189	5906340630196	cukier biały kostka n 500g	c011811_1	0