UNIWERSYTET EKONOMICZNY W POZNANIU

Wybrane algorytmy uczenia maszynowego na potrzeby klasyfikacji zawodów i specjalności

Aneta Czyżewska

Promotor: dr Maciej Beręsewicz, prof. UEP

Uniwersytet Ekonomiczny w Poznaniu

Poznań, 05.07.20221

Plan prezentacji

1. Wybrane algorytmy uczenia maszynowego na potrzeby klasyfikacji zawodów i specjalności
   - Cel pracy
   - Cele badawcze
   - Źródła danych
   - Metody wykorzystane w pracy
   - Zbudowane klasyfikatory
   - Rozkład danych
2. Wyniki
   - Wyniki klasyfikacji
3. Podsumowanie
4. Literatura

Cel pracy

Głównym celem badania było **zbudowanie klasyfikatora przepisującego 1 cyfrowe kody zawodów do treści ogłoszenia**.

al. Niepodległości 10, 61-875 Poznań
tel. +48 61 856 90 00
fax +48 61 852 57 22

Cele badawcze

C1) określenie, który algorytm spośród badanych pozwoli na najdokładniejszą klasyfikację

C2) przygotowanie zbioru uczącego i testowego zawierającego dane wejściowe w formie opisów zawodów i przypisanego 6-cio cyfrowego kodu,

C3) wytrenowanie modelu opartego na nowych danych, w celu weryfikacji założeń wynikających z modeli testowych,

Źródła danych użyte w pracy

- Centralna Baza Ofert Pracy (CBOP)
- Ręcznie kodowane dane zebrane przez Instytut Badań Edukacyjnych w Warszawie

al. Niepodległości 10, 61-875 Poznań
tel. +48 61 856 90 00
fax +48 61 852 57 22

Metody wykorzystane w pracy

- Język: Python (numpy, pandas, regex, nltk.corpus, matplotlib.pyplot, seaborn, sklearn)
- Metody wykorzystane przy budowie modeli: Pipelines, CountVectorizer, Transformator Tf-idf
- Algorytmy na podstawie których zostały zbudowane modele:
  1. Naiwny klasyfikator Bayes'a
  2. Metoda wektorów nośnych (SVM) ze Stochastycznym zejściem gradientowym (SGD)
  3. Regresja logistyczna

Zbudowane klasyfikatory

- Dla danych CBOP:
  - Naiwny klasyfikator Bayesa dla modeli wielomianowych
  - Liniowy klasyfikator SVM
  - Klasyfikacja logistyczna oparta na modelu regresji logistycznej
- Klasyfikator ze zbiorem treningowym opartym na danych z CBOP i testowym z ręcznie przypisanymi kategoriami z IBE:
  - Naiwny klasyfikator Bayesa dla modeli wielomianowych
  - Liniowy klasyfikator SVM
  - Klasyfikacja logistyczna oparta na modelu regresji logistycznej

Rozkład danych

	Rekordy z CBOP	Rekordy z IBE
1. Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy	11224	582
2. Specjaliści	75635	3152
3. Technicy i inny średni personel	65891	1346
4. Pracownicy biurowi	77780	827
5. Pracownicy usług i sprzedawcy	138888	1251
7. Robotnicy przemysłowi i rzemieślnicy	134159	1504
8. Operatorzy i monterzy maszyn i urządzeń	76059	719
9. Pracownicy wykonujący prace proste	124251	681

Tab. 1: Rozkłady danych w utworzonych modelach

Źródło: Opracowanie własne.

Wyniki klasyfikacji

Tab. 2: Wynik klasyfikacji na zbiorze CBOP

Algorytm	**NB**	**SVM**	**LogReg**
średni ważony wynik f-1	78%	79%	86%
Najlepiej przewidziane klasy	5, 4, 2	5, 7 ,2	5, 9, 4
Najgorzej przewidziane klasy	1, 3	3, 1	3

Tab. 3: Wynik klasyfikacji na zbiorze IBE

Algorytm	**NB**	**SVM**	**LogReg**
średni ważony wynik f-1	58%	68%	74%
Najlepiej przewidziane klasy	2	2, 5	2, 5, 9
Najgorzej przewidziane klasy	1, 3	1, 3	3, 1

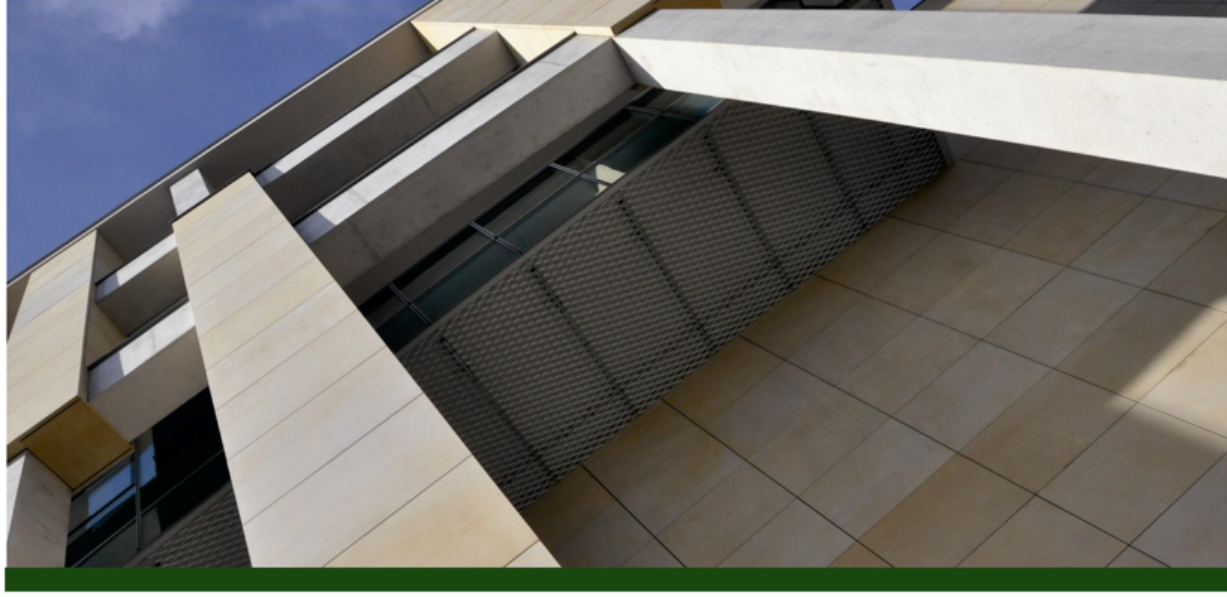
Podsumowanie

- Dla każdego zbioru danych utworzono trzy klasyfikatory w celu wyboru optymalnego algorytmu do klasyfikacji ogłoszeń o pracę
- Najlepszy okazał się model oparty na regresji logistycznej z dokładnością 74%.
- Zauważono duży wpływ rozkładu danych na wyniki predykcji modeli
- Poszczególne modele wykazały największą trudność w rozróżnieniu kategorii 7 z 9 (Pracownicy wykonujący proste prace z osobami technicznymi i innym średnim personelem), 3 z 2 (średni personel ze specjalistami) oraz 2 z 1 (specjaliści z przedstawicielami władz publicznych i urzędnikami).

Literatura

- Elias, P., Birch, M. (2010). SOC 2010: The revision of the Standard Occupational Classification 2000,
- Ministerstwo Rodziny i Polityki Społecznej. (2014). Klasyfikacja Zawodów i specjalności z 2017 roku
- Tukey, J. W. (1977). Exploratory data analysis (Vol. 2). Reading, MA.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pret- tenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Jounal of Machine Learning Research, 12, 2825–2830.

UNIWERSYTET
EKONOMICZNY
W POZNANIU



Dziękuję