Sztuczne inteligencja i inżynieria wiedzy

Laboratorium

Ćwiczenie 4. Podstawy maszynowego uczenia na przykładzie klasyfikacji tekstu

Opracowanie: Maciej Piasecki, Arkadiusz Janz

Cel ćwiczenia

Zapoznanie się z wybranymi metodami maszynowego uczenia się poprzez ich zastosowanie w zagadnieniu klasyfikacji tekstu. Skupimy się przede wszystkim na dwóch metodach omówionych na wykładzie:

- naiwnym klasyfikatorze bayesowskim (NB),
- drzewach decyzyjnych (DT).

Mają one różny charakter. Pierwsza to dość prosta, ale często dość skuteczna metoda statystycznego uczenia. Natomiast drzewa decyzyjne w bardzo przejrzysty sposób wykorzystują cechy charakteryzujące obiekty i przez to dobrze ilustrują wiele zjawisk w maszynowym uczeniu. Drzewa decyzyjne, szczególnie współcześnie, nie są metodą pierwszego wyboru w odniesieniu do klasyfikacji tekstu. Zostały jednak wybrane na potrzeby tego ćwiczenia, ponieważ stawiają wysokie wymogi przed selekcją cech i ich użycie do klasyfikacji tekstu jest pouczającym wyzwaniem dla zastosowań metod selekcji cech.
Obiektami podlegającymi klasyfikacji będą dokumenty tekstowe, a celem będzie ich przypisanie do określonych, ustalonych z góry kategorii tematycznych. Dokumenty tekstowe pozwalają na dość naturalne określenie cech opisujących je i ich wydobycie przy pomocy prostych metod.
Dokumenty tekstowe jako obiekty klasyfikacji charakteryzują się jednak bardzo dużą liczbą potencjalnych cech. Z punktu widzenia tego ćwiczenia to bardzo dobrze, ponieważ ostatnim z celów jest zapoznanie się z praktycznym działaniem metod selekcji cech, przede wszystkim tych, które zostały omówione na wykładzie.

Rozpoznawanie kategorii tekstu

Zagadnienie klasyfikacji tekstu to proces automatycznego tagowania treści lub elementów treści dokumentów tekstowych za pomocą dostępnego zestawu potencjalnych kategorii (etykiet, klas). Trudność danego problemu klasyfikacji tekstu zależy zarówno od jakości pozyskanych danych tekstowych jak również i charakteru docelowego zestawu kategorii, jakie model ma rozpoznawać. W trakcie realizacji tego ćwiczenia skoncentrujemy się na klasyfikacji pełnych dokumentów tekstowych, gdzie źródłowe teksty oraz docelowy zestaw rozważanych kategorii zaczerpnięte zostały ze struktury Wikipedii.

Wzorcowa kolekcja etykietowanych danych to podzbiór artykułów z Wikipedii z przypisanymi kategoriami Wikipedii (z dolnego poziomu). Zestaw przypisanych kategorii składa się z 34 wybranych etykiet. Dane oryginalnie podzielono na dwie części:

- część opisana jako treningowa: Wikipedia, 2015, *Wiki train - 34 categories*, CLARIN-PL digital repository, http://hdl.handle.net/11321/738
- część opisana jako testowa: Wikipedia, 2015, *Wiki test - 34 categories*, CLARIN-PL digital repository, http://hdl.handle.net/11321/739

W ramach zadania należy zaprojektować system uczenia maszynowego do rozpoznawania wskazanych kategorii Wikipedii (34 etykiety). W tym celu należy zaprojektować zestaw cech

generowanych na podstawie treści dokumentów, które będą stanowiły reprezentację tekstu, oraz program, który taki zestaw cech wygeneruje.

Uwaga: **nie należy używać żadnych metadanych, w tym elementów struktury dokumentów jako podstawy do definiowania i wygenerowania cech – należy się skupić wyłącznie na tekstowej zawartości dokumentów, czyli w pierwszej kolejności na wyrazach w tekstach.**

Na podstawie wskazanych metod klasyfikacji (NB, DT) oraz przygotowanej reprezentacji tekstu należy przygotować program do klasyfikacji tekstu, który będzie podstawą do przeprowadzenia badań skuteczności opracowanych rozwiązań. Na tym etapie należy zapoznać się z wybranym systemem lub pakietem uczenia maszynowego (sugerowane jest wykorzystanie systemu *Weka* http://www.cs.waikato.ac.nz/ml/weka/), który umożliwia selekcję cech, trenowanie klasyfikatorów, dostrajanie parametrów oraz testowanie. Dane do przeprowadzenia badań należy podzielić zgodnie z zasadami znanymi z wykładu oraz literatury. W każdym układzie k-krotna walidacja krzyżowa jest obowiązkowa.

Realizacja ćwiczenia

1. Zapoznanie się z wykładem oraz rozdziałem 13 z "Introduction to Information Retrieval" (dodatkowa literatura rozszerzająca) – opisana w bibliografii.
2. Zapoznanie się z wybranym systemem lub pakietem do maszynowego uczenia.
3. Zapoznanie się ze strukturą, zawartością i metadanymi (sposobem etykietowania) dokumentów tekstowych ze wzorcowej kolekcji (uwaga: etykiety pochodzą ze struktury Wikipedii – jaka może być ich spójność i konsekwencja w przypisaniu dokumentów do kategorii? kto przypisuje artykuły do kategorii w Wikipedii?):

- dane należy połączyć w jeden zbiór i później podzielić zgodnie z zasadami przedstawionymi na wykładzie (patrz punkt 7).

4. Zaprojektowanie zestawu cech generowanych na podstawie treści dokumentów. Przypominamy: nie należy używać żadnych metadanych, w tym elementów struktury dokumentów jako podstawy do definiowania cech – należy się skupić wyłącznie na tekstowej zawartości dokumentów.
5. Zaprojektowanie i implementacja programu do wydobywania wartości cech z dokumentów i ich zapisywania w formacie odpowiednim dla wybranego systemu lub pakietu do maszynowego uczenia (w przypadku Weki będzie to ARFF).
6. Zaprojektowanie i skonfigurowanie systemu do maszynowego uczenia obejmującego selekcję cech, dostrajanie, trenowanie klasyfikatorów oraz testowanie.
7. Podział pozyskanych danych na odpowiednie podzbiory zgodnie z zasadami znanymi, np. z wykładu czy literatury. W każdym układzie k-krotna walidacja krzyżowa jest obowiązkowa.

8. Zaplanowanie eksperymentów (ze zrozumieniem, przemyśleniem, absolutnie nie należy robić tego mechanicznie) mających na celu wnikliwe zbadanie obu algorytmów klasyfikacji w różnych ich wariantach w ramach postawionego zadania. Plan eksperymentów powinien być przedstawiony i uzasadniony w raporcie. Należy co najmniej wziąć pod uwagę:
   - wpływ parametrów właściwych poszczególnym algorytmom,
   - metody selekcji cech i ich parametry,
   - uwypuklenie słabych i mocnych stron badanych algorytmów,
   - określenie zestawów cech istotnych dla poszczególnych klas oraz co one mówią nam o poszczególnych klasach w danych,
   - działanie klasyfikatorów dla różnych podzbiorów klas decyzyjnych (wcześniej należy dobrze się zastanowić czego szukamy w tym eksperymencie),
9. Skonfigurowanie możliwie jak najlepszego systemu do klasyfikacji zadanego zbioru przy pomocy zadanych algorytmów klasyfikacji. Wykazanie jego właściwości w przekonujący i rzetelny sposób.
10. **Badania dodatkowe (na dodatkowe punkty)**:
    - zbadanie krzywej uczenia dla różnej ilości danych treningowych,
    - wywołanie efektu przeuczenia na różne sposoby i zbadanie efektów tego zjawiska.

Raport powinien obejmować opis podjętych decyzji oraz ich uzasadnienie. Ze szczególną uwagą powinny być opisane zaplanowane eksperymenty, osiągnięte rezultaty i wyciągnięte wnioski. Raport nie musi i nie powinien być za długi, a jedynie trafnie i treściwie napisany.

Zadanie dodatkowe (na dodatkowe punkty)
Zapoznanie się z wybranymi algorytmami maszynowego uczenia, selekcji cech lub reprezentacji tekstu na potrzeby klasyfikacji semantycznej, które nie były omawiane na wykładzie. Następnie zastosowanie ich do zdefiniowanego problemu i przeprowadzenie przewidzianych kroków.
Uwaga: warunkiem zaliczenia tego zadania dodatkowego jest bardzo dobre i z pełnym zrozumieniem poznanie wybranych metod dodatkowych. Mechaniczne stosowanie różnych implementacji i 'generowanie' masy rezultatów nie spotka się z uznaniem punktowym. Celem głównym ćwiczenia 4 jest nauczenie się wybranych metod.

Punktacja

<u>Etap 1</u>
2 pkt – zaprojektowanie reprezentacji danych, ich wczytanie i wygenerowanie reprezentacji w wymaganym formacie.
1 pkt – zaprojektowanie i skonfigurowanie systemu do maszynowego uczenia.
1 pkt – podział danych (i ew. ich konwersja) do przeprowadzenia badań eksperymentalnych.
Kod źródłowy systemu. Sprawozdanie z prac (2-3 stron). Specyfikacja platformy badawczej. Kolekcja danych wejściowych i wyjściowych.

Etap 2
4 pkt – Zaplanowanie i przeprowadzenie eksperymentów oraz opracowanie ich rezultatów. Sprawozdanie z prac (4-6 stron, lub więcej). Kolekcje danych wejściowych i wyjściowych. Rysunki, diagramy, wykresy.

Etap 3
2 pkt – Skonfigurowanie możliwie jak najlepszego systemu do klasyfikacji zadanego zbioru przy pomocy zadanych algorytmów klasyfikacji. Specyfikacje parametrów. Kolekcje danych wejściowych i wyjściowych. Rysunki, diagramy, wykresy.
Sprawozdanie z prac (1-4 stron).

Zadanie dodatkowe do 5 pkt:

- realizacja badań dodatkowych
- realizacja zadania dodatkowego.

Sprawozdanie z prac (5-6 stron). Specyfikacje parametrów. Kolekcje danych wejściowych i wyjściowych. Rysunki, diagramy, wykresy.

Bibliography

1. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. *Introduction* to. *Information*. *Retrieval*. Cambridge University Press, 2008. (there will be also a copy in the Board):
http://www-nlp.stanford.edu/IR-book
lub
https://archive.org/details/AnIntroductionToInformationRetrieval
lub
http://www-connex.lip6.fr/~gallinar/livres%20-%20fichiers/2007-%20Manning-irbookonlinereading.pdf
2. Weka documentation: http://www.cs.waikato.ac.nz/ml/weka/documentation.html
3. Artykuły zasugerowane w Weka dla wybranych algorytmów maszynowego uczenia.
4. Paweł Cichosz. Systemy uczące się. Wyd. NT, Warszawa, 2000.