Identyfikacja dokumentów tekstowych

1 Cel laboratorium

Zadaniem do wykonania jest text mining, gdzie zamiast tabeli z atrybutami stosuje się dokumenty tekstowe.

W zadaniach będą użyte zestawy danych składające się się z setek artykułów prasowych. Każdy artykuł został sklasyfikowany z wartością 1, jeżeli artykuł jest związany z tytułem dokumentu lub z wartością 0, jeśli artykuł nie ma związku. Celem zadania jest sprawdzenie, jak WEKA klasyfikuje artykuły opierając się wyłącznie na słowach w artykule. W tym celu każde słowo sprowadzi się do rangi atrybutu. I tak albo atrybut będzie binarny wskazując czy słowo znajduje się w dokumencie, albo będzie liczbą całkowitą określającą ile razy słowo pojawia się w dokumencie (liczba słów).

2 Opis danych

Oba zbiory danych zawierają zbiór artykułów prasowych na różne tematy. Pliki zostały przetworzone na format ARFF.

Dane można znaleźć pod adresem `http://www.sc.ehu.es/ccwbayes/docencia/mmcc/files/UCI-datasets/`

1. ReutersCorn-train.arff zawiera 1554 artykułów, z których 45 dotyczy kukurydzy.
2. ReutersGrain-train.arff zawiera 1554 artykułów, w tym 103 o ziarnie.

3 Wykorzystanie atrybutów binarnych

Aby użyć słów z każdego artykułu jako atrybuty,trzeba skorzystać z filtru StringToWordVector (weka.filters.unsupervised.attribute.StringToWordVector).

3.1 Zadanie 1: Klasyfikacja artykułów na podstawie słów zawartych w artykule

Należy załadować plik ReutersCorn-train.arff.

1. Wybierz filtr: StringToWordVector i zmień następujące ustawienia: lowerCaseTokens = True, wordsToKeep = 2500. Wybierz AlphabeticTokenizer jak tokenizer. Zastosuj filtr i przejrzyj listę atrybutów. Co ten filtr robi? Jaki procent przypadków jest dodatni (klasy o etykiecie 1)?

2. Przejdź do panelu klasyfikacji. Sklasyfikuj dane za pomocą: Naive Bayes i SMO. Zanotuj TP i FP dla klasy 1 i obszar pod krzywą ROC.

3. Porównaj wyniki z powyższych klasyfikatorów. Który model jest lepszy w klasyfikacji i dlaczego (oprzyj spostrzeżenia na wartościach TP / FP i powierzchni ROC)? Znaczenie TP / FP mogą się zmianiać w zależności od celu klasyfikacji i wykorzystanych danych. Na przykład, jeśli klasyfikujemy spam od dobrych wiadomości, to jednym z najważniejszych celów będzie utrzymanie FP bliskie zeru, ponieważ w przeciwnym razie model jest skłonny wyrzucać dobre e-maile. Wysoki poziom TP jest również dobry, ale w tej sytuacji lepiej jest mieć kilka wiadomości spamowych przepuszczonych przez filtr, niż tracić ważne maile. Co według Ciebie jest ważniejsze w klasyfikacji artykułów prasowych, TP czy FP i dlaczego?

4. Dokonaj ponownej klasyfikacji z użyciem Attribute Selected Classifier w porównaniu z powyższymi klasyfikatorami. Zastosuj InfoGain jako estymator atrybutów i Ranker jako metodę wyszukiwania. W ustawieniach Ranker'a, zmień 2 w parametrze numToSelect na 100. Oznacza to, że algorytm wyszuka wszystkie atrybuty, ale wykorzysta do klasyfikacji tylko 100 najlepszych. Zanotować wartości TP i FP oraz obszar ROC każdego modelu.

5. Zobaczyć wyniki Ranker'a. Przejrzyj listę 100 słów (atrybutów), które zostały wybrane. Jak osądzasz ten wybór? Czy słowa są związane z kukurydzą? Przeanalizuj 20 pierwszych słów i zanotuj wszelkie, które Twoim zdaniem są nieistotne (lub nie są specyficznae dla artykułów o kukurydzy). Wyjaśnij, dlaczego znalazły się na liście.

6. Porównać wyniki z klasyfikatów. Czy ranking atrybutów daje lepsze czy gorsze wyniki. Dlaczego?

3.2 Zadanie 2: Klasyfikacja artykułów na podstawie częstości występowania słów

Należy załadować plik ReutersGrain-train.arff.

1. Wybierz filtr: StringToWordVector i zmień następujące ustawienia: lowerCaseTokens = True, wordCount = True, wordsToKeep = 2500. Wybierz AlphabeticTokenizer jak tokenizer. Zastosuj filtr i przejrzyj listę atrybutów. Jaka jest różnica w wartościch atrybutów i co te liczby oznaczają?

2. Przejdź do panelu klasyfikacji. Sklasyfikuj dane za pomocą: Naive Bayes Multinomia. Zanotuj TP i FP dla klasy 1 i obszar pod krzywą ROC.

3. Porównaj wyniki z powyższych klasyfikatorów. Który model jest lepszy w klasyfikacji i dlaczego (oprzyj spostrzeżenia na wartościach TP / FP i powierzchni ROC)?

4. Wróć do panelu Preprocess i cofnij filtr StringToWordVector. Wykonaj ponownie, ale zmień ustawienie WordCount na false. Wróć do panelu klasyfikacji i ponownie sklasyfikuj dane za pomocą Attribute Selected Clas- sifier z klasyfikatorem Naive Bayes. Zastosuj Infogain jako jako estymator atrybutów i Ranker jako metodę wyszukiwania. Uruchom test trzy razy, przy pierwszym uruchomieniu ustawić numToSelect (w ustawieniach Ranker'a) na 100, za drugim razem zmienić go na 50 i za trzecim razem zmienić na 25. Zanotuj TP / FP i obszar ROC dla każdego testu.

5. Porównać wyniki z klasyfikatów. Zanotuj swoje odkrycia i uzasadnienia.

6. Porównaj wyniki z bieżącego zadania z wynikami z klasyfikacji binarnej. Czy SMO jest skuteczniejsze z atrybutami binarnymi czy liczbą słów? Które wyniki są lepsze Naive Bayes czy Naive Bayes Multinomial. Wyjaśnij dlaczego tak jest.

3.3 Sprawozdanie

Zanotowane odpowiedzi i wnioski wysłać na adres jkolodziejczyk[at]wi.zut. edu.pl w formacie pdf w pliku o nazwie *imie_ nazwisko.pdf* Tytuł maila: Sprawozdanie z ZSIwMET. Opóźnienia będą wpływały na obniżenie punktacji. Najkorzystniej byłoby wykonać zajęcia na laboratoriach i wysłać od razu wyniki.