Informatyka, studia dzienne, II st. semestr II

Komputerowe systemy rozpoznawania **2013/2014**
Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 10:30

Data oddania: _____ Ocena: _____

Mateusz Grotek 186816
Paweł Tarasiuk 186875

Zadanie 1.: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem niniejszego zadania jest zbadanie różnych metod ekstrakcji cech oraz miar podobieństwa dla tekstu i zastosowanie ich w procesie klasyfikacji słabym klasyfikatorem kNN. Omówione zostaną metody znane w literaturze oraz nasze własne pomysły, a wszystkie te elementy znajdą swoje odzwierciedlenie w przygotowanej przez nas implementacji, co umożliwi nam wygenerowanie i ocenę wyników działania różnych metod.

2. Wprowadzenie

Klasyfikator kNN (k nearest neighbours, k najbliższych sąsiadów) to słaby klasyfikator, posiadający jednak istotną zaletę jaką jest prostota implementacji. Dlatego jest on szeroko stosowany, gdyż pomimo swojej prostoty daje relatywnie dobre wyniki. Dodatkowo można go użyć w technikach typu AdaBoost, które pozwalają z kilku słabych klasyfikatorów stworzyć jeden silny. Jego idea polega na wybraniu takiej kategorii która jest najczęstsza wśród sąsiadów wektora podlegającego klasyfikacji. Potencjalni sąsiedzi są dostarczani do klasyfikatora na etapie jego uczenia. Aby wskazać najbliższych sąsiadów należy oczywiście użyć jakiejś miary odległości, czyli metryki. Istnieje też odmiana klasyfikatora, którą moglibyśmy określić jako k najbardziej podobnych sąsiadów. W tej odmianie zamiast odległości w zadanej metryce używana jest funkcja podobieństwa. Pokazuje to jak szerokie zastosowanie może mieć ten klasyfikator.

W naszym projekcie użyliśmy następujących funkcji podobieństwa:
— miara Jaccarda,
— metoda n-gramów,
— autorska metoda bazująca na słowach kluczowych
Użyte zostały następujące metryki:
— metryka euklidesowa
— metryka uliczna
— metryka Czebyszewa

Klasyfikacji można dokonywać bezpośrednio na podanych tekstach, jednak takie rozwiązanie charakteryzuje się wysokim kosztem obliczeniowym. Dlatego często lepiej jest wyekstrachować ze zbiorów pewne wektory cech, które pozwolą szybko sklasyfikować dane teksty. Wybór wektorów cech zależy od postawionego zadania. W naszym projekcie użyliśmy trzech sposobów ekstrakcji cech:
— naiwny sposób bazujący na użyciu wszystkich słów
— bazujący na macierzy częstości terminów
— autorska matoda bazująca na zbiorach rozmytych

Sama klasyfikacja została wykonana na dwóch zbiorach danych, każdy podzielony na podzbiór uczący i podzbiór testowy, przy czym sposób podziału jest wybierany w aplikacji. Można dokonać podziału w sposób losowy (przy zadanych proporcjach), lub w ustalony sposób (60% początkowych tekstów jako zbiór uczący, reszta to zbiór testowy). Pierwszy zbiór danych, to zestaw krótkich wiadomości prasowych firmy Reuters. Drugi zestaw przygotowany przez nas, to teksty znanych pisarzy. Teksty prasowe zostały sklasyfikowane na podstawie dwóch kategorii: miejsca którego dotyczą i tematu. Jeśli chodzi o teksty znanych pisarzy, to zostały one sklasyfikowane kategorią, którą są nazwiska pisarzy.

Poniżej omówimy autorskie metody ekstrakcji cech i miarę podobieństwa, których użyliśmy. Główną ideą naszej metody ekstrakcji są słowa kluczowe. Zbiór takich słów należy przygotować dla konkretnego zadania. Może on być wyznaczony ręcznie, lub wygenerowany automatycznie na podstawie danych uczących. Zbiór takich słów dzielimy na podzbiory na podstawie ich znaczenia. Można oczywiście używać tylko jednego podzbioru w którym są umieszczone wszystkie słowa kluczowe, ale zwiększenie ilości podzbiorów może poprawić klasyfikację, ze względu na zwiększenie wymiarowości przestrzeni. W jednym podzbiorze powinny być słowa oznaczająca podobne obiekty, na przykład nazwy geograficzne. Następnie dla każdej kategorii względem której klasyfikujemy przygotowujemy osobny zestaw podzbiorów. Inaczej mówiąc jeżeli kategorią jest „kanada" a jednym w jednym z podzbiorów znajdują się „organizacje", to tworzymy podzbiór „kanadyjskie organizacje". Widać, że jest to metoda analogiczna do tego jak ludzie kategoryzują obiekty. Wystarczy następnie zauważyć, że określenie „kanadyjskie organizacji" możemy modelować jako pewien zbiór rozmyty. Niektóre organizacje są stricte kanadyjskie, inne tylko częściowo. Na podstawie danych uczących, lub wiedzy eksperckiej jesteśmy w stanie podać funkcję przynależności do danego zbioru.

Mając taki zestaw zbiorów możemy wyekstrachować cechy z tekstów zbioru uczącego i na tej podstawie nauczyć klasyfikator. Ekstrakcja przebiega następująco. Wstępnie każdy element wektora cech jest liczbą, która jest sumą

wartości funkcji przynależności dla każdego słowa, kluczowego z podanego zbioru rozmytego, które wystąpiło w podanym tekście. Opisuje to wzór

$$f_1(T,n) = \sum_{s \in K(n) \cap T} \mu_{K(n)}(s),$$

przy czym T to tekst z którego ekstrachujemy cechy, n to numer składowej wektora cech którą obliczamy, $K(n)$ to n-ty zbiór rozmyty, $\mu(s)$ to funkcja przynależności dla zbioru $K(n)$. Następnie wektor taki jest normalizowany do długości równej 1. Przy obliczaniu funkcji podobieństwa również użyliśmy słów kluczowych. Dla wybranych słów kluczowych liczony jest prosty współczynnik dopasowania (simple matching coefficient).

3. Opis implementacji

Implementacja programu została wykonana w języku Java.

Należy tu zamieścić krótki i zwięzły opis zaprojektowanych klas oraz powiązań między nimi. Powinien się tu również znaleźć diagram UML (diagram klas) prezentujący najistotniejsze elementy stworzonej aplikacji. Należy także podać, w jakim języku programowania została stworzona aplikacja.

4. Materiały i metody

W tym miejscu należy opisać, jak przeprowadzone zostały wszystkie badania, których wyniki i dyskusja zamieszczane są w dalszych sekcjach. Opis ten powinien być na tyle dokładny, aby osoba czytająca go potrafiła wszystkie przeprowadzone badania samodzielnie powtórzyć w celu zweryfikowania ich poprawności (a zatem m.in. należy zamieścić tu opis architektury sieci, wartości współczynników użytych w kolejnych eksperymentach, sposób inicjalizacji wag, metodę uczenia itp. oraz informacje o danych, na których prowadzone były badania). Przy opisie należy odwoływać się i stosować do opisanych w sekcji drugiej wzorów i oznaczeń, a także w jasny sposób opisać cel konkretnego testu. Najlepiej byłoby wyraźnie wyszczególnić (ponumerować) poszczególne eksperymenty tak, aby łatwo było się do nich odwoływać dalej.

5. Wyniki

W tej sekcji należy zaprezentować, dla każdego przeprowadzonego eksperymentu, kompletny zestaw wyników w postaci tabel, wykresów itp. Powinny być one tak ponazywane, aby było wiadomo, do czego się odnoszą. Wszystkie tabele i wykresy należy oczywiście opisać (opisać co jest na osiach, w kolumnach itd.) stosując się do przyjętych wcześniej oznaczeń. Nie należy tu komentować i interpretować wyników, gdyż miejsce na to jest w kolejnej sekcji. Tu również dobrze jest wprowadzić oznaczenia (tabel, wykresów) aby móc się do nich odwoływać poniżej.

6. Dyskusja

Sekcja ta powinna zawierać dokładną interpretację uzyskanych wyników eksperymentów wraz ze szczegółowymi wnioskami z nich płynącymi. Najcenniejsze są, rzecz jasna, wnioski o charakterze uniwersalnym, które mogą być istotne przy innych, podobnych zadaniach. Należy również omówić i wyjaśnić wszystkie napotakane problemy (jeśli takie były). Każdy wniosek powinien mieć poparcie we wcześniej przeprowadzonych eksperymentach (odwołania do konkretnych wyników). Jest to jedna z najważniejszych sekcji tego sprawozdania, gdyż prezentuje poziom zrozumienia badanego problemu.

7. Wnioski

W tej, przedostatniej, sekcji należy zamieścić podsumowanie najważniejszych wniosków z sekcji poprzedniej. Najlepiej jest je po prostu wypunktować. Znów, tak jak poprzednio, najistotniejsze są wnioski o charakterze uniwersalnym.

Literatura

Na końcu należy obowiązkowo podać cytowaną w sprawozdaniu literaturę, z której grupa korzystała w trakcie prac nad zadaniem (przykład na końcu szablonu)

8. Wersja wstępna algorytmu