Zaawansowane metody i techniki analizy danych

OPIS PROJEKTU

Celem projektu jest zastosowanie nabytej wiedzy do samodzielnie przeprowadzonej analizy różnorodnych danych oraz wyciągania uogólnionych wniosków na ich podstawie.

Termin oddania projektu podaje prowadzący i jest on nieprzekraczalny!

Cześć 1 Data Minning- WEKA

1. Znajdź odpowiednie dane do analizy. Dane muszą zawierać kilkaset przykładów oraz atrybuty ciągłe i dyskretne. Możesz skorzystać z przykładowych zbiorów ze strony: http://www.cs.waikato.ac.nz/ml/weka/. Dokonaj wstępnego przygotowania danych np. wypełnij brakujące wartości atrybutów, dokonaj dyskretyzacji atrybutów ciągłych.

2. Przeprowadź badania dla dwóch różnych algorytmów klasyfikacji. Jeśli to możliwe uwzględnij różne ustawienia parametrów algorytmów, spróbuj określić ich wpływ na wyniki. Oceń wyniki uczenia się za pomocą walidacji krzyżowej.

3. Przeprowadź badania dla dwóch różnych algorytmów grupowania. Jeśli to możliwe uwzględnij różne ustawienia parametrów algorytmów, spróbuj określić ich wpływ na wyniki. Oceń wyniki grupowania ze względu na liczbę grup. Znajdź/wygeneruj (mały) zbiór danych, w których znany jest podział na grupy. Sprawdź, który z wybranych algorytmów generuje podział na klastry zbliżony do prawdziwego.

4. Na podstawie wybranego zbioru wygeneruj 3 zbiory o różnych wielkościach np. 5000, 5000, 10000 poprzez powielenie przykładów. Przeprowadź eksperyment stosując algorytm Apriori. Porównaj liczbę powstałych reguł asocjacyjnych i czas ich tworzenia dla zbiorów danych o różnych rozmiarach oraz dla różnych wartości progu minimalnego wsparcia. Uwzględnij różne ustawienia parametrów używanego algorytmu i podejmij próbę określenia ich wpływu na wyniki.

5. Uzyskane wyniki przedstaw w czytelnej formie tabel i/lub wykresów. Przeanalizuj uzyskane wyniki, podejmij próbę ich wyjaśnienia, przedyskutuj wynikające z nich wnioski. Do sprawozdania dołącz opis uruchamianych klas i ich parametry.

Część 2 Analiza statystyczna -program R lub Matlab (do wyboru)

1. Znajdź odpowiednie do analizy dane, w którym możemy wyróżnić przynajmniej 3 populacje np. cena akcji różnych spółek przez 3 dni, skuteczność przynajmniej 3 leków testowana na pewnej grupie osób itp. Wielkość każdej populacji jest dowolna, pamiętaj jednak, że wielkość populacji może determinować wybór właściwego testu statystycznego

2. Określ parametry rozkładu poszczególnych populacji: średnia, odchylenie standardowe, minimum, maksimum, mediana, pierwszy kwartyl i trzeci kwartyl.

3. Sprawdź równość badanej cechy we wszystkich populacjach odpowiednim testem. Pamiętaj o sprawdzeniu czy zmienne są powiązane, równości wariancji, rozkładu badanej cechy. Jeśli jest potrzeba wykonaj odpowiednią analizę post-hoc.

4. Sprawdź, w której populacji badana cecha ma największą średnią/medianę. Wykonaj kilka testów parami: populacja 1 z 2, 1 z 3, 2 z 3. Wybierz do tego zadania odpowiedni test, pamiętając o sprawdzeniu czy zmienne są powiązane, równości wariancji, rozkładu badanej cechy.

5. Sprawdź odpowiednim testem czy średnia/mediana w każdej populacji jest mniejsza niż średnia/mediana+10%*średnia/mediana wyznaczona w pkt. 2.

6. Wyznacz (jeśli to możliwe lub uzasadnij dlaczego nie można) 95% i 99% przedział ufności dla średniej oraz wariancji dla każdej populacji

7. Uzyskane wyniki przedstaw w czytelnej formie tabel i/lub wykresów. Postaw właściwe hipotezy, przeanalizuj uzyskane wyniki, podejmij próbę ich wyjaśnienia, przedyskutuj wynikające z nich wnioski, uzasadnij wybór testów. Do sprawozdania dołącz ciąg wywołań funkcji.