Pozyskiwanie wiedzy
Projekt na zaliczenie laboratorium

Adam Zagdański, Artur Suchwałko

2 marca 2015

Spis treści

1 Cel projektu

- Głównym celem projektu jest zastosowanie poznanych metod pozyskiwania wiedzy (data mining) do przeprowadzenia kompletnej analizy wybranych danych, związanych z określonym zagadnieniem praktycznym.
- Ważnym elementem projektu powinno być porównanie skuteczności wykorzystywanych metod/algorytmów oraz szczegółowe wnioski dotyczące ich praktycznej przydatności, w kontekście analizowanego zagadnienia.
- Analiza powinna obejmować następujące elementy
  1. analiza opisowa + wizualizacja danych (eksploracyjna analiza danych),
  2. klasyfikacja wraz z oceną dokładności,
  3. analiza skupień wraz z oceną jakości,
  4. zastosowanie wybranej metody redukcji wymiaru w połączeniu z klasyfikacją i analizą skupień.
- Zbiory danych do wyboru wymienione są w Rozdziale 2. W uzasadnionych przypadkach i po konsultacji z prowadzącym laboratorium możliwy jest wybór danych spoza tej listy.
- Szczegóły dotyczące zawartości i struktury projektu przedstawione są w Rozdziale 3.
- Dodatkowe informacje dotyczące metod i algorytmów można znaleźć w Rozdziale 4.
- W Rozdziałach 5 umieszczono informacje o ważnych terminach.
- W Rozdziale 6 znajdują się dodatkowe uwagi dotyczące projektu.

2 Zbiory danych

Do analizy w ramach projektu wybieramy **jeden** zbiór danych z poniższej listy:

- **Diagnostyka medyczna:** Breast Cancer Wisconsin (Original) Data Set
  http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
  Warto spojrzeć również na: Breast Cancer Wisconsin (Diagnostic) Data Set,
  http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- **Diagnostyka medyczna:** Hepatitis Data Set
  http://archive.ics.uci.edu/ml/datasets/Hepatitis
- **Ryzyko kredytowe:** Statlog (German Credit Data) Data Set
  http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
- **Analizy socjoekonomiczne:** Adult Data Set
  http://archive.ics.uci.edu/ml/datasets/Adult
- **Automobile Data Set**
  http://archive.ics.uci.edu/ml/datasets/Automobile
- **Proteomika/technologia SELDI-TOF:** Arcene Data Set
  http://archive.ics.uci.edu/ml/datasets/Arcene.
- **Diagnostyka medyczna/dane mikromacierzowe:**
  http://stat.ethz.ch/~dettling/bagboost.html
  Do wyboru jeden ze zbiorów: Leukemia data, Colon data, Prostate data, Lymphoma data, SRBCT data, Brain data.
- **Filtrowanie spamu**
  http://archive.ics.uci.edu/ml/datasets/Spambase
  Uwaga: dane te są dostępne także w pakiecie R: spam{ElemStatLearn}.
- **Analiza attrition/churn (problem odchodzenia klientów)**
  http://www.dataminingconsultant.com/data/churn.txt
  Opis danych można znaleźć m.in. w książce T.Larose, Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych, PWN (2006)
- **Modelowanie odpowiedzi na kampanię marketingową:** Dane *clothing store*
  http://www.dataminingconsultant.com/data/Clothing_Store
  Opis danych można znaleźć m.in. w książce T.Larose, Metody i modele eksploracji danych, PWN (2008)
- **Marketing bankowy**
  http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Uwaga: Niektóre dane zawierają bardzo dużo przypadków i mogą wystąpić problemy związane ze złożonością obliczeniową lub efektywnością metod. Wówczas do analiz można wybrać losowo podzbiór (np. 1000 losowo wybranych obiektów). W przypadku wyboru podzbioru, proszę pamiętać, żeby opisać to szczegółowo w pracy!

3 Zawartość i struktura projektu

Lista **obowiązkowych elementów**, które powinny znaleźć się w projekcie

- **Opis problemu i sformułowanie pytań badawczych**
  Krótka informacja o specyfice zagadnienia. Na jakie pytania będziemy chcieli odpowiedzieć analizując dane? Jakie potencjalne korzyści mogą wynikać z przeprowadzonych analiz? (np. korzyść to: lepsza metoda diagnostyczna, lepsza skuteczność w wykrywaniu złych/dobrych klientów starających się o kredyt, wyodrębnienie grup klientów, którym można zaoferować określoną ofertę, identyfikacja istotnych cech/zmiennych, itp.)

- **Charakterystyka danych**
  Rozmiar danych, liczba przypadków i cech, rodzaje cech, informacja o brakujących obserwacjach, informacje o nietypowych wartościach (np. niestandardowe kodowanie brakujących obserwacji, itp.)

- **Wykorzystane metody**
  Jakie metody/algorytmy zostały wykorzystane? Do realizacji jakich zadań wykorzystano te metody/algorytmy, np.: analiza wstępna i wizualizacja (eksploracyjna analiza danych), klasyfikacja, predykcja, analiza skupień.

- **Rezultaty**
  Wyniki w formie tabel, wykresów i diagramów.
  Uwaga: proszę zamieszczać tylko najważniejsze wyniki! Pozostałe rezultaty można umieścić jako załączniki (np. plik pdf z rysunkami).

- **Wnioski**
  Precyzyjnie sformułowane wnioski: co wynika z przeprowadzonych analiz? Jak można wykorzystać wnioski w praktyce? (np. opracowanie nowej/lepszej strategii w instytucji/firmie, nowej/lepszej metody diagnostycznej, itp.)

- **Dalsze możliwości badań**
  Krótka informacja o dalszych możliwych kierunkach badań (co można/warto by jeszcze zbadać i z wykorzystaniem jakich metod?)

Zalecamy aby struktura projektu była zgodna ze **standardem IMRAD** (Introduction, Methods, Results And Discussion), patrz np. http://en.wikipedia.org/wiki/IMRAD.

4 Metody i algorytmy

- W projekcie powinny być wykorzystane wybrane metody z określonej grupy
  - **Analiza opisowa i wizualizacja**
    - Cel
      - Podstawowa charakterystyka zmiennych (m.in.: zakres wartości, własności rozkładu, itp.)
      - Analiza zależności (korelacji) cech
      - Wstępna ocena zdolności cech do dyskryminacji (separacji) obiektów
      - Identyfikacja obserwacji brakujących i odstających (nietypowych)
    - Metody/narzędzia
      - Wskaźniki sumaryczne (miary położenia i rozrzutu, wyznaczone dla wszystkich danych i w grupach/klasach)
      - Podstawowe wykresy (histogramy, wykresy rozrzutu, box-ploty, itp.)
  - **Klasyfikacja**
    - Cel: budowa reguły klasyfikacyjnej (reguły decyzyjnej)
    - Wybrane metody/algorytmy, w tym: liniowa i kwadratowa analiza dyskryminacyjna (LDA, QDA), metoda k-najbliższych sąsiadów (k-NN), inne.
    - Ocena dokładności klasyfikacji
      - Wersja podstawowa – porównanie błędu klasyfikacji z wykorzystaniem podziału na zbiór uczący i testowy dla różnych kombinacji cech i różnych metod,
      - Wersja rozszerzona – zastosowanie schematu typu *cross-validation* lub *bootstrap*.
  - **Analiza skupień**
    - Cel: pogrupowanie obiektów ze względu na występujące podobieństwo.
    - Wybrane metody/algorytmy, w tym: k-means, PAM, AGNES, inne.
    - Ocena jakości analizy skupień
      - wersja podstawowa – porównanie średnich wartości indeksu *silhouette* dla różnej liczby skupień K,
      - wersja rozszerzona – inne wskaźniki oceniające m.in. separację, zwartość skupień, itp. (w tym własne pomysły!)
  - **Redukcja wymiaru**
    - Cel: ekstrakcja cech, wizualizacja danych wielowymiarowych
    - Zastosowanie wybranych metod (np. PCA lub MDS) w powiązaniu z klasyfikacją i analizą skupień.

- **Kila dodatkowych uwag:**
  - Proszę nie stosować metod całkowicie automatycznie! Należy upewnić się, np. dla jakiego rodzaju cech można daną metodę zastosować i ewentualnie poprzedzić jej zastosowanie wyborem odpowiednich cech.
  - Do analizy skupień nie powinniśmy stosować zmiennej grupującej (etykietki klas), aby umożliwić algorytmowi odkrycie prawdziwej struktury danych!
  - Zachęcamy także do włączenia do projektu dodatkowych metod, np.: niestandardowych algorytmów klasyfikacji i analizy skupień, metod odkrywania reguł asocjacyjnych, metod wyboru najlepszych cech (*feature selection*), itp.
  - Kolejne etapy analizy nie powinny być traktowane jako całkowicie niezależne części. Warto np. zbadać czy wyniki analizy skupień potwierdzają podział na klasy zadany przez określoną zmienną grupującą, itp.

5 Ważne terminy

- **Wybór tematu projektu:** informacja powinna być przekazana prowadzącemu na zajęciach lub wysłana e-mailem najpóźniej do **31.03.2015**,
- **Punkt kontrolny:** weryfikacja postępów w pracy nad projektem, odbędzie się w połowie semestru (konkretna data będzie ustalona na zajęciach),
- **Końcowa wersja projektu** – powinna być wysłana (wraz z R-kodami) najpóźniej do **10.06.2015**,
- **Uwaga:** Brak systematyczności w pracy nad projektem (patrz: punkt kontrolny) czy też spóźnienie w oddaniu projektu będzie wiązało się z obniżeniem oceny!

6 Pozostałe uwagi

- Projekt może być realizowany w grupach (maksymalnie dwuosobowych). W przypadku realizacji projektu przez dwie osoby analiza powinna być jednak odpowiednio szczegółowa i nie może ograniczać się tylko do metod z podstawowego zakresu.
- Do projektu należy dołączyć (nie wklejać!) wykorzystany skrypt(y) w języku R.
- Po wczytaniu danych w R proszę sprawdzić czy wszystkie typy zmiennych zostały prawidłowo rozpoznane (w szczególności zmienne typu *numeric* i *factor*)
- Zakres projektu jest sformułowany dość ogólnie. Ważne jest zaprezentowanie umiejętności precyzyjnego postawienia problemu i wykorzystania do jego analizy odpowiednich technik uczenia statystycznego/data mining.
- Zachęcamy również do wykazania się własną inwencją! W razie wątpliwości czy określona analiza mieści się w ramach projektu można skonsultować pomysł z prowadzącym laboratorium.
- Przykłady niestandardowych analiz
  - Uwzględnienie macierzy kosztów w ocenie jakości klasyfikacji
  - Zagadnienia związane z wyborem cech, np.:
    - Przedziałowanie cech ciągłych w połączeniu z analizą efektywności metod (np. zamiast ciągłej zmiennej DOCHÓD, wykorzystujemy zmienną nominalną DOCHÓD.PRZEDZ przyjmującą wartości {'mały','średni','duży'})
    - Konstrukcja cech pochodnych w połączeniu z analizą efektywności metod (np. zamiast WIEK i PŁEĆ wykorzystujemy cechę produktową WIEK×PŁEĆ)
- Oryginalność analiz i niestandardowe pomysły będą premiowane!