Przygotowanie danych do klasyfikacji

1 Cel laboratoriów

Poznanie procesu przygotowania danych do uczenia klasyfikatora. Aby dobrze zaobserwować różnice przeprowadzane będą testy przed i po zastosowaniu pewnego kroku przygotowania wstępnego danych.

2 Na czym polega wstępne przygotowanie danych (Data preprocessing and transformation)

2.1 Przygotowanie danych

Na tym etapie głównie skupiamy się na rozwiązaniu dwóch problemów:

1. poradzić sobie z danymi zaszumionymi
2. poradzić sobie z danymi brakującymi

I tak czyszczenie danych z szumu polega na:

- odnalezieniu rekordów powtarzających się
- odnalezieniu niewłaściwych wartości atrybutów
- smoothing data (wygładzenie danych)

Jeżeli w zbiorze danych pojawiają się rekordy, w których brakuje danych można wykonać jeden ze schematów radzenia sobie z takimi danymi:

- usunąć rekordy z tymi danymi
- uzupełnić średnią z wartości atrybutu rekordów w tej samej klasie
- uzupełnić średnią z wartości atrybutu rekordów najbardziej podobnych

2.2 Transformacja danych

Na tym etapie dokonuje się przekształcenia danych polegających na:

- normalizacji
- konwersji typów
- wyborze atrybutów i rekordów

3 Dane do wykorzystania

Korzystając z `http://archive.ics.uci.edu/ml/datasets/Credit+Approval` UCI Repository of Machine Learning Databases pobrać zbiór danych archiwalnych dotyczących przydzielania kart kredytowych (Credit Approval Data Set).

4 Zadania do wykonania

1. Pobrać dane z repozytorium i stworzyć z nich plik w formacie arff (Plik załączyć jako rozliczenie wykonanego zadania — wysłać w mailu ze sprawozdaniem).
2. Wczytać przygotowany plik i przejrzeć statystyki o atrybutach. Zanotować w sprawozdaniu to co uznano, za istotne. Jeżeli coś jest w danych ciekawe napisać o tym!
3. Wykonać na zbiorze klasyfikację z użyciem klasyfikatora J48 w klasie tree (klasyfikator implementuje algorytm C4.5) `http://en.wikipedia.org/wiki/C4.5_algorithm`. W sprawozdaniu zawrzeć wyniki z eksperymentu.
4. Przejrzeć dane pod kątem powtarzających się rekordów i błędnych wartości atrybutów.
   (a) Zaraportować w sprawozdaniu, jeżeli istnieją takie same rekordy. Usunąć ze zbioru takie rekordy, o ile występują. Jeżeli usunięto rekordy ponownie dokonać klasyfikacji J48 i odnotować wyniki klasyfikacji.
   (b) Zastosować fitr *NumericCleaner*. Krótko opisać na czym transformacja polega i co się po jego zastosowaniu zmieniło się w danych. Dokonać klasyfikacji J48. Zapisać wnioski.
5. Przejrzeć dane pod kątem brakujących wartości:
   (a) Usunąć rekordy z brakującymi wartościami. Dokonać klasyfikacji J48 i zapisać wyniki.
   (b) Zastosować filtry WEKA:
      i. *ReplaceMissingValues*
      ii. *EMImputation*

      Dokonać klasyfikacji J48 i zapisać wyniki.

Porównać podejścia i skonstruować wnioski.

6. Dokonać różnego rodzaju transformacji (w sprawozdaniu krótko opisać na czym transformacja polega i co się po jej zastosowaniu zmieniło w danych) na danych i dokonać klasyfikacji J48. Porównać wyniki i zapisać wnioski.
   (a) *Normalize*
   (b) *NominalToBinary*
   (c) *Discretize (supervised)*
   (d) *PKIDiscretize*
   (e) *RemoveUseless*
   (f) *Attribute Selection*

4.1 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_nazwisko.pdf* należy przesłać w terminie do 5go listopada na adres jkolodziejczyk[at]wi.zut.edu.pl. Tytuł maila: Sprawozdanie 1 PZMSI.