



Wstępne przetwarzanie danych

Oracle Data Mining

Uwaga!

- W trakcie ćwiczeń powstaną tabele: `bin_cat`, `bin_num`, `norm_lin`, `clip`.
- Należy je usunąć przed powtórzeniem ćwiczeń.

```
DROP TABLE bin_cat;  
DROP TABLE bin_num;  
DROP TABLE norm_lin;  
DROP TABLE clip;
```

Przygotowanie

- Utwórz perspektywę v_pracownicy w której znajduje się rok zatrudnienia pracownika zamiast daty.

```
CREATE VIEW V_PRACOWNICY AS
SELECT
  id_prac, imie,
  nazwisko, etat,
  id_szefa, EXTRACT (YEAR FROM zatrudniony)
  AS rok_zatrudnienia,
  placa_pod, placa_dod, id_zesp
FROM pracownicy;
```


Dyskretyzacja atrybutów nominalnych (1)

- Parametr N – liczba binów
- Wybieranych jest N najczęściej/najrzadziej występujących wartości. Tworzą one N binów.
- Pozostałe wartości trafiają do $N+1$ -szego bina

Dyskretyzacja atrybutów nominalnych (2)

- Utworzenie tabeli do zapisywania listy otrzymanych binów.

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_CAT(
    bin_table_name=>'bin_cat', ← Nazwa tabeli (dowolna)
    bin_schema_name=>'infXXXXX_XX'); ← Nazwa schematu (opcjonalna)
END;
/
```

Dyskretyzacja atrybutów nominalnych (3)

- Utworzenie binów (działa domyślnie na wszystkich kolumnach typu łańcuchowego)

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_CAT_FREQ(
  bin_table_name=>'bin_cat',
  data_table_name=>'v_pracownicy',
  bin_num=>3,
  exclude_list=>DBMS_DATA_MINING_TRANSFORM.
  column_List('NAZWISKO','IMIE'));
END;
/
```

Nazwa analizowanej tabeli

Nazwa tabeli wynikowej

Liczba binów

Których kolumn nie analizować

Dyskretyzacja atrybutów nominalnych (4)

- Utworzenie perspektywy z wynikami dyskretyzacji:

```
BEGIN
```

```
  DBMS_DATA_MINING_TRANSFORM.XFORM_BIN_CAT(
```

```
    bin_table_name=>'bin_cat',
```

Nazwa tabeli z binami

```
    data_table_name=>'v_pracownicy',
```

```
    xform_view_name=>'v_bin_cat_prac');
```

Nazwa
dyskretyzowanej
tabeli

```
END;
```

```
/
```

Nazwa
wynikowej
perspektywy

Ćwiczenie

- Obejrzyj strukturę tabeli DMDATA.AUTOS
- Wykonaj dyskretyzację atrybutu MAKE w tabeli DMDATA.AUTOS na 5 binów
- Obejrzyj listę binów w tabeli bin_cat
- Napisz zapytanie do tabeli DMDATA.AUTOS, które wyświetla wszystkie marki samochodów posortowane malejąco wg. liczby egzemplarzy. Porównaj wynik zapytania z otrzymanymi binami.
- Utwórz i obejrzyj perspektywę ze zdyskretyzowanym atrybutem.

Dyskretyzacja atrybutów numerycznych (1)

- Utworzenie tabeli do zapisywania listy otrzymanych binów.

```
BEGIN  
DBMS_DATA_MINING_TRANSFORM.CREATE_BIN_NUM(  
    bin_table_name=>'bin_num',  
    bin_schema_name=>'infXXXXX_XX');  
END;  
/
```

Nazwa tabeli (dowolna)

Nazwa schematu
(opcjonalna)

Dyskretyzacja atrybutów numerycznych (2)

- Utworzenie binów (dyskretyzacja equalWidth, działa na wszystkich atrybutach typu liczbowego)

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_NUM_EQWIDTH(
  bin_table_name=>'bin_num',
  data_table_name=>'v_pracownicy',
  bin_num=>4,
  exclude_list=>DBMS_DATA_MINING_TRANSFORM.
    column_List('ID_PRAC',
      'ID_SZEFA', 'PLACA_POD',
      'PLACA_DOD', 'ID_ZESP'),
  round_num=>4);
END;
```

Nazwa
analizowanej
tabeli

Nazwa tabeli
wynikowej

Liczba binów

Których kolumn nie
analizować

Dyskretyzacja atrybutów numerycznych (3)

- Utworzenie binów (dyskretyzacja equalFrequency, działa na wszystkich atrybutach typu liczbowego)

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.INSERT_BIN_NUM_QTILE(
    bin_table_name=>'bin_num',
    data_table_name=>'v_pracownicy',
    bin_num=>3,
    exclude_list=>DBMS_DATA_MINING_TRANSFORM.
      Column_List('ID_PRAC', 'ID_SZEFA',
        'ROK_ZATRUDNIENIA', 'PLACA_DOD',
        'ID_ZESP' ));
END;
/
```

Nazwa
analizowanej
tabeli

Nazwa tabeli
wynikowej

Liczba binów

Których kolumn nie
analizować

Dyskretyzacja atrybutów numerycznych (4)

- Utworzenie perspektywy z wynikami dyskretyzacji:

```
BEGIN
```

```
  DBMS_DATA_MINING_TRANSFORM.XFORM_BIN_NUM(  
    bin_table_name=>'bin_num',  
    data_table_name=>'v_pracownicy',  
    xform_view_name=>'v_bin_num_prac');
```

```
END;
```

```
/
```

Nazwa tabeli z binami

Nazwa
dyskretyzowanej
tabeli

Nazwa
wynikowej
perspektywy

Ćwiczenie

- Wykonaj dyskretyzację płacy podstawowej pracowników na 3 kategorie o jednakowej szerokości.
- Utwórz perspektywę z wynikiem dyskretyzacji.
- Obejrzyj wyniki.

Normalizacja atrybutów (1)

- Utworzenie tabeli pomocniczej.

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.CREATE_NORM_LIN(
  norm_table_name=>'norm_lin', ← Nazwa tabeli (dowolna)
  norm_schema_name=>'*****'); ← Nazwa schematu (opcjonalna)
END;
```

Normalizacja atrybutów (2)

- Normalizacja metodą min-max

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_MINMAX(
    norm_table_name=>'norm_lin',
    data_table_name=>'v_pracownicy',
    exclude_list=>DBMS_DATA_MINING_TRANSFORM.
      Column_List('ID_SZEFA', 'ROK_ZATRUDNIENIA',
        'PLACA_POD', 'PLACA_DOD', 'ID_ZESP'));
END;
/
```

Nazwa
tabeli z
danymi

Nazwa tabeli
pomocniczej

Których kolumn nie
przetwarzać

Normalizacja atrybutów (3)

- Standaryzacja

```
BEGIN
```

```
  DBMS_DATA_MINING_TRANSFORM.INSERT_NORM_LIN_ZSCORE(  
    norm_table_name=>'norm_lin',  
    data_table_name=>'v_pracownicy',  
    exclude_list=>DBMS_DATA_MINING_TRANSFORM.  
      Column_List('ID_PRAC', 'ROK_ZATRUDNIENIA',  
        'PLACA_POD', 'PLACA_DOD', 'ID_ZESP'));
```

Nazwa
tabeli z
danymi

Nazwa tabeli
pomocniczej

Których kolumn nie
przetwarzać

```
END;  
/
```

Normalizacja atrybutów (4)

- Utworzenie tabeli z wynikami normalizacji

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_NORM_LIN(
    norm_table_name=>'norm_lin',
    data_table_name=>'v_pracownicy',
    xform_view_name=>'v_norm_lin_pracownicy');
END;
/
```

Nazwa
tabeli z
danymi

Nazwa tabeli
pomocniczej

Nazwa perspektywy z
wynikiem normalizacji

Ćwiczenie

- Dokonaj normalizacji atrybutu CITY_MPG metodą Min-Max, i standaryzacji atrybutu HIGHWAY_MPG w tabeli DMDATA.AUTOS.
- Utwórz perspektywę z wynikami normalizacji.
- Obejrzyj wynik

Przycinanie atrybutów (1)

- Utworzenie tablicy pomocniczej

```
BEGIN
```

```
  DBMS_DATA_MINING_TRANSFORM.CREATE_CLIP(
```

```
    clip_table_name=>'clip',
```

```
    clip_schema_name=>'*****');
```

```
END;
```

Nazwa tabeli (dowolna)

Nazwa schematu
(opcjonalna)

Przycinanie atrybutów (2)

- Zastąpienie X% największych i najmniejszych wartości wartością NULL

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.
  INSERT_CLIP_TRIM_TAIL(
    clip_table_name=>'clip',
    data_table_name=>'v_pracownicy',
    tail_frac=>0.3,
    exclude_list=>DBMS_DATA_MINING_TRANSFORM.
      Column_List('ID_PRAC', 'ROK_ZATRUDNIENIA',
        'PLACA_POD', 'PLACA_DOD', 'ID_SZEFA'));
END;
/
```

Nazwa tabeli
pomocniczej

Nazwa tabeli z danymi

X% w postaci
ułamka

Które
kolumny
zignorować

Przycinanie atrybutów (3)

- Zastąpienie X% największych i najmniejszych wartości wartościami granicznymi

```
BEGIN
DBMS_DATA_MINING_TRANSFORM.
  INSERT_CLIP_WINSOR_TAIL(
    clip_table_name=>'clip',
    data_table_name=>'v_pracownicy',
    tail_frac=>0.3,
    exclude_list=>DBMS_DATA_MINING_TRANSFORM.
      Column_List('ID_PRAC', 'ID_SZEFA',
        'ROK_ZATRUDNIENIA', 'PLACA_POD',
        'ID_ZESP' ));
END;
/
```

X% w postaci
ułamka

Nazwa tabeli
pomocniczej

Nazwa tabeli z danymi

Które
kolumny
zignorować

Przycinanie atrybutów (4)

- Utworzenie perspektywy z danymi po przycięciu:

```
BEGIN
  DBMS_DATA_MINING_TRANSFORM.XFORM_CLIP(
    clip_table_name=>'clip',
    data_table_name=>'v_pracownicy',
    xform_view_name=>'v_clip_pracownicy');
END;
/
```

Nazwa
tabeli z
danymi

Nazwa tabeli
pomocniczej

Nazwa perspektywy z
wynikiem przycięcia

Ćwiczenie

- Przytnij atrybuty numeryczne: usuń 20% wartości atrybutu CITY_MPG, zastępując usuwane wartości wartościami brzegowymi, obetnij także 20% wartości atrybutu HIGHWAY_MPG zastępując ucięte wartości wartością NULL.
- Utwórz perspektywę z wynikami przycięcia
- Obejrzyj uzyskany wynik.