CES: A Real World Dataset for Association Rule Mining

Eduardo Corrêa Gonçalves
egoncalves@ic.uff.br

1. Introduction

The **CES Dataset** is a real database that contains observations collected from a household survey called "Consumer Expenditure Survey". This survey has been conducted by a Brazilian institute of research since 1947 to, among other goals, support the analysis of food consumption of Brazilian families.

The database keeps data about 1540 interviewed families from eight distinct Brazilian cities. The dataset stores the list of products acquired by each family on their last visit to a supermarket as well as demographic data of these families (city of residence, number of members in the family and monthly income).

The dataset is provided in the CSV and ARFF formats, which are briefly described in the next sections.

2. CSV Format - "ces_hybrid.csv"

In this format, the first column contains the transaction ID (each ID represents a distinct family) and the second contains one item. The columns are separated by comma. Each transaction has the following structure:

- **First item**: City of residence. List of possible values:
  - "City_Belem";
  - "City_Belo_Horizonte";
  - "City_Curitiba";
  - "City_Florianopolis";
  - "City_Fortaleza";
  - "City_Goiania";
  - "City_Porto Alegre";
  - "City_Recife".

- **Second item**: Monthly income of the family. List of possible values:
  - "Income_below_2.5" (up to 2.5 minimum salaries);
  - "Income_2.5_to_5" (between 2.5 and 5 minimum salaries);
  - "Income_5_to_8" (between 5 and 8 minimum salaries);
  - "Income_8_to_12" (between 8 and 12 minimum salaries);
  - "Income_12_to_18" (between 12 and 18 minimum salaries);
  - "Income_18_to_25" (between 18 and 25 minimum salaries);
  - "Income_25_to_43" (between 25 and 43 minimum salaries);
  - "Income_above_43" (more than 43 minimum salaries);

- **Third item**: Number of members in the family. List of possible values:
    - “Members_1” (one member, i.e., a person who lives alone);
    - “Members_2” (family with two members);
    - “Members_3” (family with three members);
    - “Members_4” (family with four members);
    - “Members_5” (family with five members);
    - “Members_6” (family with six members);
    - “Members_above_6” (family with seven or more members).

- **Remainder items**: the remainder items associated to the transaction will store the products acquired by the family.

Example:

```
40017,City_Curitiba
40017,Income_above_43
40017,Members_5
40017,banana
40017,black_beans
40017,chocolate_truffle
40017,potato
40017,vinegar
40017,yeast
50161,City_Florianopolis
50161,Income_5_to_8
50161,Members_3
50161,cucumber
50161,egg
50161,watermelon
```

The above example shows the demographic information and the list of items purchased by two distinct families, identified by the transactions ID’s 40017 and 50161, respectively.

Observe that the family identified by “40017” lives in the city of Curitiba (first item), with monthly income above 43 minimum salaries (second item) and is composed by 5 members (third item). This family purchased six products: “banana”, “black beans”, “chocolate truffle”, “potato”, “vinegar” and “yeast”.

Analogously, the family identified by “50161” lives in Florianopolis, with monthly income between 5 and 8 minimum salaries and is composed by 3 members, having purchased three products: “cucumber”, “egg”, and “watermelon”.

The CSV format is suitable for use with the R “arules” package. The below script presents a basic example (in this script, consider that the dataset “ces_hybrid.csv” is stored in the “c:\tmp” folder).

```
#imports the arules package

library("arules")


#imports the CES dataset

ces_td <- read.transactions("c:\\tmp\\ces_hybrid.csv", format = "single", cols = c(1,2), 
sep=",")


#generates the association rules

#minimum support = 3%; minimum confidence = 60%; maximum length = 2

rules <- apriori(ces_td, parameter = list(support = 0.03, confidence = 0.6, maxlen=2))


#save the results

write(rules, file = "c:\\tmp\\results.csv", sep = ",", col.names = NA)
```

3. ARFF Format - “ces_hybrid.arff”

ARFF is the format adopted by the well-known Weka data mining tool. All items (including demographic information) are declared as binary variables in the header section. The transactions are structured in the sparse format, where items with value 0 (i.e., items that do not belong to the transaction) do not need to be explicitly represented.

Example:

```
{0 1,11 1,21 1,66 1,93 1,94 1, 749 1}
```

In this example, the transaction is composed by the following items: 0, 11, 21, 66, 93, 94 and 749. Item 0 corresponds to the first item declared in the header section (“City_Belem”). Analogously, 11 corresponds to the 12nd item declared in the header section (“Members_4”). And so on.

4. Citing the CES Dataset

If you want to refer to the “CES dataset” in a publication, please cite the following paper:

Gonçalves, E. C. (2014). A Human-Centered Approach for Mining Hybrid-Dimensional Association Rules. Proceedings of the 17th International Conference on Information Fusion, (FUSION 2014), Salamanca, Spain.