Assigment on association rules

Rafał Rysiejko, 423827

09/02/2020

Introduction

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. Association Rules find all sets of items (itemsets) that have support greater than the minimum support and then using the large itemsets to generate the desired rules that have confidence greater than the minimum confidence. The lift of a rule is the ratio of the observed support to that expected if X and Y were independent. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design, and store layout.

In the following project, I will try to demonstrate the use of association rule learning on moderately sized demographics dataset. This data set consists of N = 8993 questionnaires filled out by shopping mall customers in the San Francisco Bay Area[1] (Impact Resources, Inc., Columbus OH, 1987).

Performing association rules analysis on such demographic data might help to uncover relationships that could be leveraged by marketers to direct marketing efforts more effectively by utilizing socio-economic dependencies.

For the purposes of the analysis, we will use the first 14 questions relating to demographics and socio-economic information. These questions are listed in Table 1. The data are seen to consist of a mixture of ordinal and (unordered) categorical variables. The complete questionnaire with answers is added in appendix 1.

Table 1: Variable description

Variable	Demographic	#Values	Type
1	Sex	2	Caterogical
2	Marital status	5	Caterogical
3	Age	7	Ordinal
4	Education	6	Ordinal
5	Occupation	9	Caterogical
6	Income	9	Ordinal
7	Years in Bay Arean	5	Ordinal
8	Dual Incomes	3	Caterogical
9	Number in household	9	Ordinal
10	Number of children	9	Ordinal
11	Household status	3	Caterogical
12	Type of home	5	Caterogical
13	Ethnic classification	8	Caterogical
14	Language at home	3	Caterogical

Explanatory data analysis

Installing and running the libraries:

```
requiredPackages = c("tidyverse","stats","ggforce","knitr",
                     "kableExtra","DataExplorer","reshape2",
```

[1]Source: Impact Resources, Inc., Columbus, OH (1987). Relevant information available here: link

```
                    "arules","arulesViz","arulesCBA","caret",
                    "arulesSequences","psych","pastecs","fastDummies")

for(i in requiredPackages){if(!require(i,character.only = TRUE)) install.packages(i)}
for(i in requiredPackages){library(i,character.only = TRUE)}
```

Loading the data:

```
data_full <- read.csv("dataset.csv",stringsAsFactors = F)
```

As a first step, I investigated what percentage of *NA* values in each analyzed variable. The results are shown in Table 2.

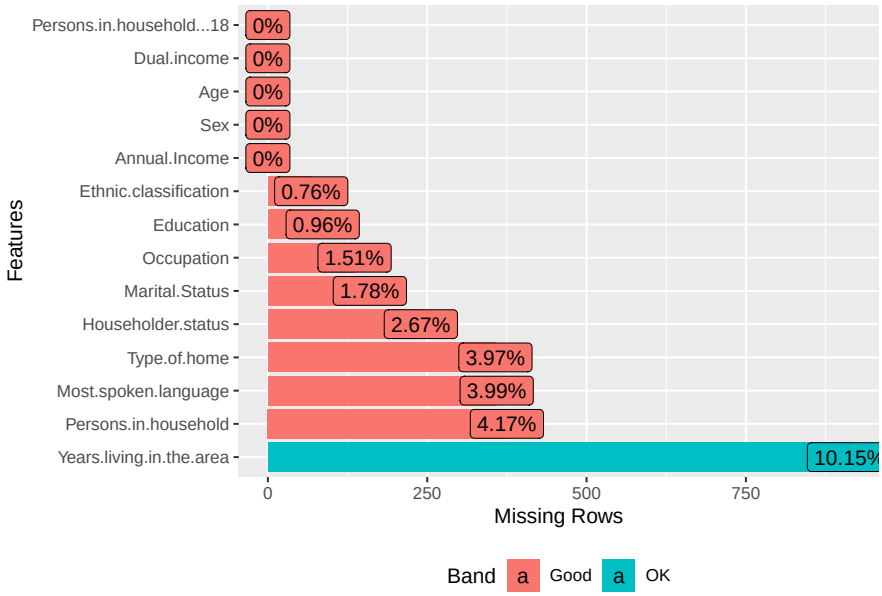

Figure 1: Genre distribution.

The most problematic variable is `Years in Bay Area` with more than 10% of missing values. As the association rule learning algorithms can not deal with *NA* values, I removed them from the dataset. This resulted in 6876 observations, which is 76.5 % of the original dataset.

```
data <- data_full[complete.cases(data_full),]
```

Next, I analyzed the descriptive statistics of the remaining observations of nominal variables in the dataset. The results are displayed in Table 2.

Then I took a deeper look into the distributions of ordinal and categorical variables.

Table 2: Summary statistics of the dataset

	nbr.val	nbr.na	min	max	range	median	mean
Annual.Income	6876	0	1	9	8	6	5.05
Sex	6876	0	1	2	1	2	1.55
Marital.Status	6876	0	1	5	4	3	3.00
Age	6876	0	1	7	6	3	3.41
Education	6876	0	1	6	5	4	3.89
Occupation	6876	0	1	9	8	3	3.64
Years.living.in.the.area	6876	0	1	5	4	5	4.21
Dual.income	6876	0	1	3	2	1	1.55
Persons.in.household	6876	0	1	9	8	3	2.86
Persons.in.household...18	6876	0	0	9	9	0	0.69
Householder.status	6876	0	1	3	2	2	1.83
Type.of.home	6876	0	1	5	4	1	1.82
Ethnic.classification	6876	0	1	8	7	7	6.01
Most.spoken.language	6876	0	1	3	2	1	1.12

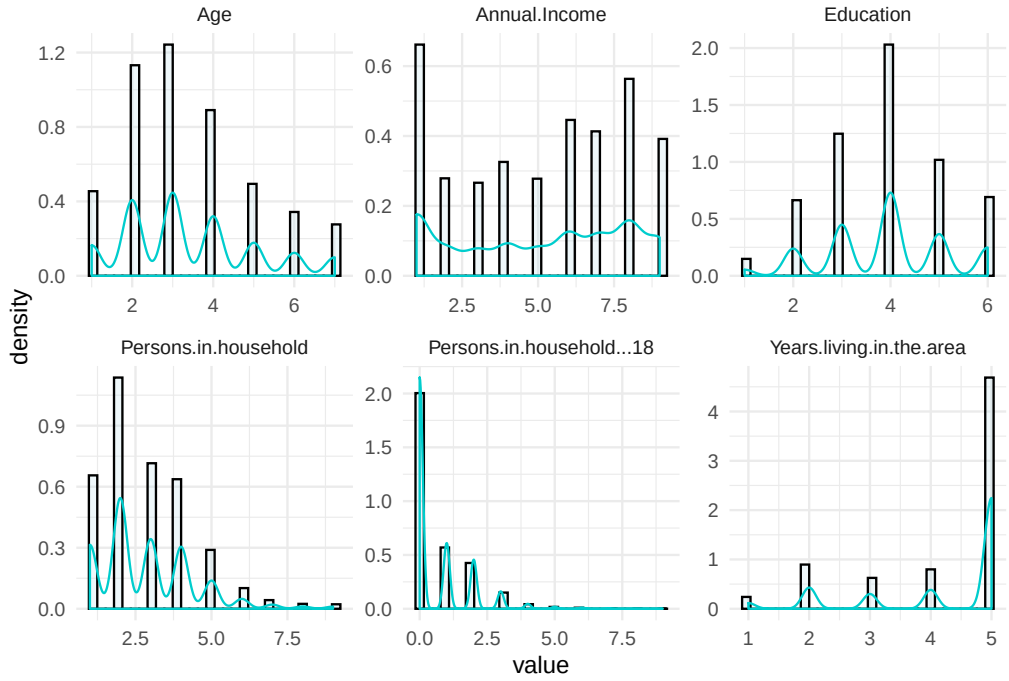

Figure 2: Histograms of ordinal variables.

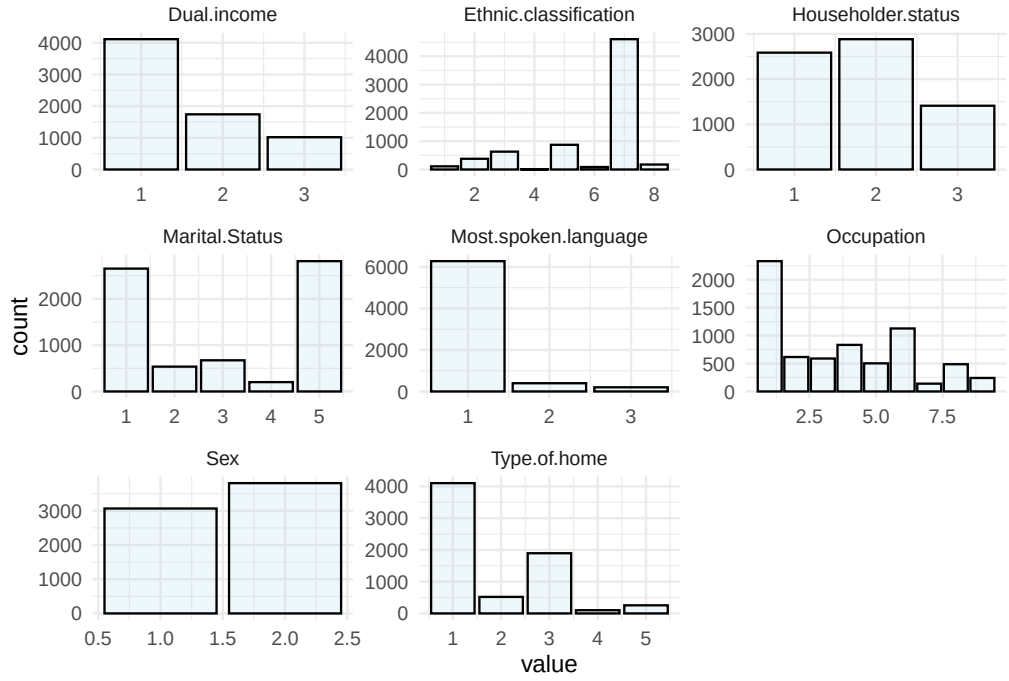

Figure 3: Barplots of categorical variables.

After removing observations with missing values, each ordinal predictor was cut at its median and coded by two dummy variables; each categorical predictor with k categories was coded by I dummy variables. This resulted in a 6876 $\times$ 50 matrix of 6876 observations on 50 dummy variables.

```
# Transofrming variables

# Categorical
categorical <- c("Sex","Marital.Status","Occupation","Dual.income","Householder.status",
                 "Type.of.home","Ethnic.classification","Most.spoken.language")

data_edit <- data

data_edit$Sex <- as.factor(data_edit$Sex)
data_edit$Marital.Status <- as.factor(data_edit$Marital.Status)
data_edit$Occupation <- as.factor(data_edit$Occupation)
data_edit$Dual.income <- as.factor(data_edit$Dual.income)
data_edit$Householder.status <- as.factor(data_edit$Householder.status)
data_edit$Type.of.home <- as.factor(data_edit$Type.of.home)
data_edit$Ethnic.classification <- as.factor(data_edit$Ethnic.classification)
data_edit$Most.spoken.language <- as.factor(data_edit$Most.spoken.language)

results <- fastDummies::dummy_cols(data_edit,
                                   select_columns = categorical,remove_selected_columns = T)

#Ordinal data
ordinal <- c("Annual.Income","Age","Education","Years.living.in.the.area",
             "Persons.in.household","Persons.in.household...18")

data_edit2 <- results
data_edit2$Annual.Income <- ifelse(data_edit2$Annual.Income > median(data_edit2$Annual.Income),1,0)
```

```
data_edit2$Age <- ifelse(data_edit2$Age > median(data_edit2$Age),1,0)
data_edit2$Education <- ifelse(data_edit2$Education > median(data_edit2$Education),1,0)
data_edit2$Years.living.in.the.area <- ifelse(data_edit2$Years.living.in.the.area >= median(data_edit2$
data_edit2$Persons.in.household <- ifelse(data_edit2$Persons.in.household > median(data_edit2$Persons.i
data_edit2$Persons.in.household...18 <- ifelse(data_edit2$Persons.in.household...18 > median(data_edit2

data_edit2 <- fastDummies::dummy_cols(data_edit2,
                                      select_columns = ordinal,remove_selected_columns = T)
data_edit2 <- as.data.frame(data_edit2)

## Association rule learning

data_edit3 <- data_edit2

data_edit3 <- apply(data_edit2, 2, as.logical)
trans <- as(data_edit3, "transactions")
```

Association rule learning

Figure 4. represents the relative frequency of each dummy variable in data. We can see that some of them are overrepresented, for example, `Most.Spoken.Language_1` (English) while others such as occupation are under-represented, except the first and fifth level. Those prevalent categories are highly likely to appear more often in the rules.

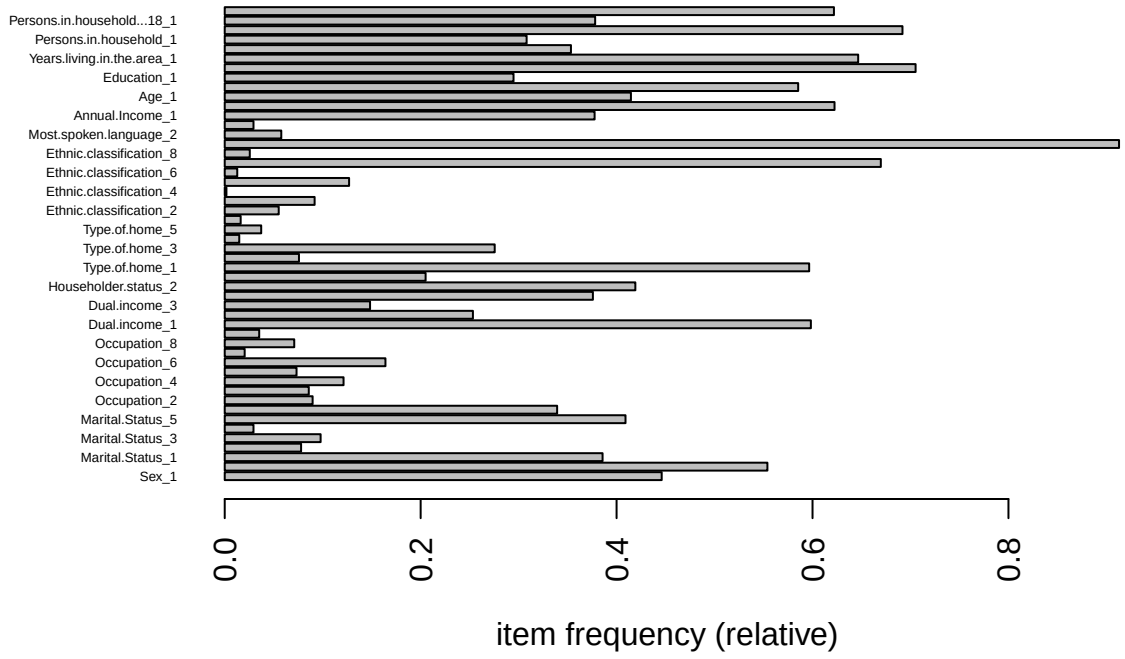

Figure 4: Item Frequency.

```
rules.trans<-apriori(trans, parameter=list(supp=0.2, conf=0.8,minlen=2))
rules.by.conf<-sort(rules.trans, by="confidence", decreasing=TRUE)

rules.by.lift<-sort(rules.trans, by="lift", decreasing=TRUE)

rules.by.count<- sort(rules.trans, by="count", decreasing=TRUE)
```

```
rules.by.supp<-sort(rules.trans, by="support", decreasing=TRUE)
```

The apriori algorithm found a total of 713 association rules with the support of at least 20% and confidence of at least 80%. Above are listed some examples of found association rules.

Association rule 1: Support 31%, confidence 99.1%, lift 1.08.

$$\begin{bmatrix} Ethnic\ classification = White \\ Age >= 35 \\ Years\ living\ in\ the\ area\ >= 7 \end{bmatrix}$$

$$Then$$

$$Most\ spoken\ language = English$$

Association rule 2: Support 20.1%, confidence 99.1%, lift 1.47.

$$\begin{bmatrix} Householder\ status = Own \\ Marital\ status = Married \\ Years\ living\ in\ the\ area\ >= 7 \end{bmatrix}$$

$$Then$$

$$Age >= 35$$

Association rule 3: Support 24.6%, confidence 83.8%, lift 1.67

$$\begin{bmatrix} Householder\ status = Own \\ Education >= College \end{bmatrix}$$

$$Then$$

$$Income >= 40,000;\ USD$$

This can be interpreted: 24.6% of asked people own their houses and have at least college education. 83.8 % of those people also have an annual income higher than 40,000 USD. Lift, on the other hand, represents the 67% increase in the expectation that someone will have an annual income higher than 40,000 USD when we know that they own their houses and have at least college education. This is a conditional probability.

We might also find results related to given items, for example, if we were interested only in finding associations with the high-income category. This might be useful for targeted marketing.

```
rules.trans.high_income<-apriori(trans, parameter=list(supp=0.1, conf=0.5,minlen=2),
                                 appearance = list(default="lhs",rhs="Annual.Income_1"),
                                 control = list (verbose=F))
rules.by.conf.high_income<-sort(rules.trans.high_income, by="confidence", decreasing=TRUE)
```

Association rule 4: Support 10.4%, confidence 87.5%, lift 2.31.

$$\begin{bmatrix} Marital\ status = Married \\ Occupation = Professional/Managerial \\ Householder\ status = Own \end{bmatrix}$$

$$Then$$

$$Income >= 40,000\ USD$$

To better illustrate association rules in the analyzed dataset, I used association rules charts from the package `arulezVIZ`

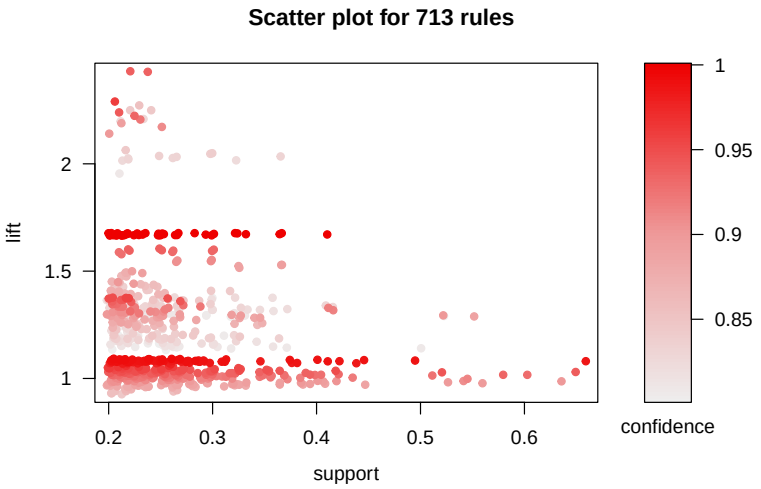

Figure 5: Item Frequency.

Figure 5 is a Two - key plot with three important metrics: lift, confidence, and support. We can see that rules with high confidence have a lower lift as well as an inverse relationship between lift and support.

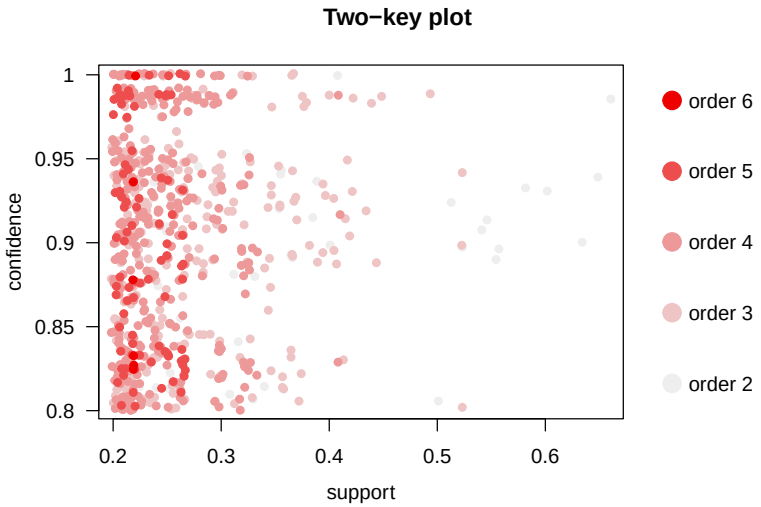

Figure 6: Item Frequency.

Figure 6 is another Two - key plot with an emphasis on "order," i.e., the number of items contained in the rule. From the plot, it is clear that order and support have an inverse relationship, which is a known fact for association rules (Seno and Karypis 2005)[2]

[2]Seno M, Karypis G (2005). "Finding Frequent Itemsets Using Length-Decreasing Support Constraint." Data Mining and Knowledge Discovery, 10, 197–228.

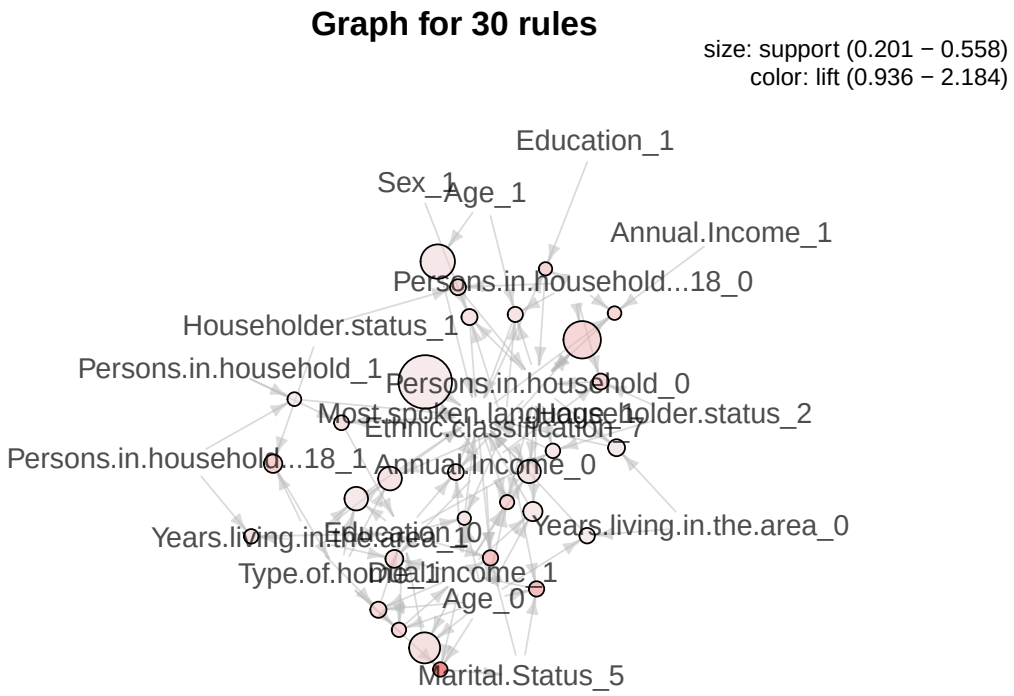

Figure 7: Item Frequency.

Figure 7 represents items and rules as vertices connecting them with directed edges. This representation focuses on how the rules are composed of individual items and shows which rules share items. The sample was needed to reduce to 30 because for larger rule sets visual analysis becomes difficult since with an increasing number of rules also the number of crossovers between the lines increases.

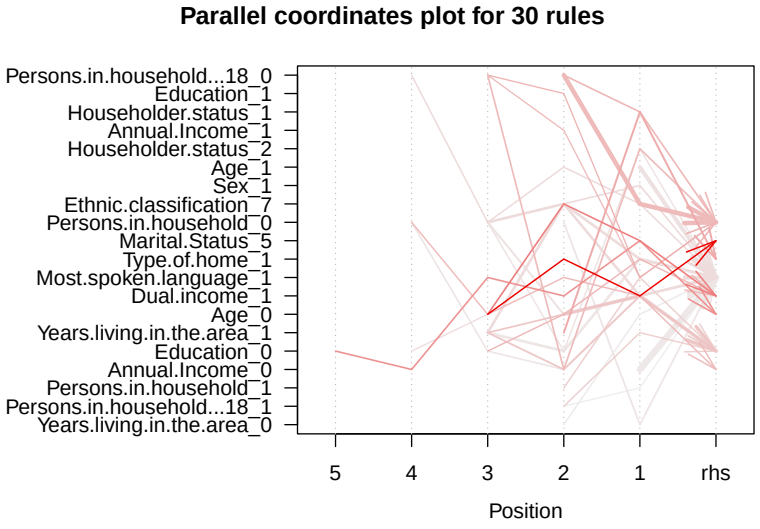

Figure 8: Item Frequency.

Figure 8 shows a parallel coordinates plot for 30 rules. The width of the arrows represents support, and the intensity of the color represents confidence.

Conclusions

Association rule is a powerful tool for data mining. The typical for association rule *market basket analysis* can also be applied to demographical data helping to uncover relationships between items from huge databases, providing insights that might be leveraged in marketing, sales, or recommendation systems.