Clustering Association Rules Mining Problem: a Symbolic Approach

Pietro Sala
Department of Computer Science
University of Verona, Italy
pietro.sala@univr.it

Data Mining & Knowledge Discovery
Academic Year 2022/2023
Exam Assignment 1

Given a universe of item $\mathtt{It}$, a linear transaction is a function $\mathtt{ltr} : \mathtt{It} \rightarrow \mathbb{N}$. The the set of all possible transactions is $\mathbb{LTr} = \mathbb{N}^{|\mathtt{It}|}$. A Linear Transaction Dataset (LTD) is a **finite** a multiset $\mathtt{LTd} : \mathbb{LTr} \rightarrow \mathbb{N}$, moreover let $|\mathtt{LTd}| = \sum_{\mathtt{ltr} \in \mathbb{LTr}} \mathtt{LTd}(\mathtt{ltr})$.

Let $\mathbb{I}_{\mathbb{N}}$ the set $\mathbb{I}_{\mathbb{N}} = \{[n, n'] \in \mathbb{N}^2 : n \leq n'\}$ the set of all the possible intervals on $\mathbb{N}$, and let $\sqsubseteq$ the relation on $\mathbb{I}_{\mathbb{N}}^2$ such that $[n, n'] \sqsubseteq [\overline{n}, \overline{n}']$ if and only if $n \leq \overline{n} \leq \overline{n}' \leq n'$. Moreover, given $\overline{n} \in \mathbb{N}$ we say that $\overline{n} \in [n, n']$ if and only if $n \leq \overline{n} \leq n'$.

A Clustering Itemset is a function $\mathtt{X}_{\mathcal{C}} : \mathtt{It} \rightarrow \mathbb{I}_{\mathbb{N}} \cup \{[0, +\infty)\}$, moreover for each $\mathtt{it} \in \mathtt{It}$, moreover given a $\mathtt{X}_{\mathcal{C}}$ and an item $\mathtt{it} \in \mathtt{It}$ we say that $\mathtt{it}$ is unconstrained by $\mathtt{X}_{\mathcal{C}}$ if and only if $\mathtt{X}_{\mathcal{C}}(\mathtt{it}) = [0, +\infty)$ (if $\mathtt{X}_{\mathcal{C}}(\mathtt{it}) \neq =[0, +\infty)$ we will say that $\mathtt{it}$ is unconstrained).

Given an LTD $\mathtt{LTd}$ and a $\mathtt{X}_{\mathcal{C}}$ we define the support (on $\mathtt{LTd}$) of $\mathtt{X}_{\mathcal{C}}$ as:

$$\mathcal{S}up(\mathtt{X}_{\mathcal{C}}) = \frac{\displaystyle\sum_{\substack{\mathtt{ltr} \in \mathbb{LTr} \text{ such that:} \\ \text{for each } \mathtt{it} \in \mathtt{It} \\ \mathtt{ltr}(\mathtt{it}) \in \mathtt{X}_{\mathcal{C}}(\mathtt{it})}} \mathtt{LTd}(\mathtt{ltr})}{|\mathtt{LTd}|}$$

Let us notice that all the concepts of frequent itemsets as well as any measures like confidence may be rephrased in terms of the newly defined support for clustering items.

A Clustering Association Rule is a rule of the form:

$$\mathtt{X}_{\mathcal{C}} \longrightarrow \mathtt{Y}_{\mathcal{C}}$$

where:

- there exist $\mathtt{it}, \mathtt{it}' \in \mathtt{It}$ such that $\mathtt{X}_{\mathcal{C}}(\mathtt{it}) \neq [0, +\infty)$ and $\mathtt{Y}_{\mathcal{C}}(\mathtt{it}') \neq [0, +\infty)$;
- for each $\mathtt{it} \in \mathtt{It}$ we have that $\mathtt{X}_{\mathcal{C}}(\mathtt{it}) \neq [0, +\infty)$ implies $\mathtt{Y}_{\mathcal{C}}(\mathtt{it}) = [0, +\infty)$, and $\mathtt{Y}_{\mathcal{C}}(\mathtt{it}) \neq [0, +\infty)$ implies $\mathtt{X}_{\mathcal{C}}(\mathtt{it}) = [0, +\infty)$.

Analogously of what has been done for standard association rules we can define the set of frequent clustering itemsets $\mathbb{Fi}_{\mathcal{C}}$ according to a given threshold $0 \leq \epsilon \leq 1$.

The Clustering Association Rules Mining Problem (CARM) consists of determining, given an LTD $\mathtt{LTd}$ and two thresholds $0 \leq \epsilon, \delta \leq 1$, the set of all and only the rules $\mathtt{X}_{\mathcal{C}} \longrightarrow \mathtt{Y}_{\mathcal{C}}$ such that $\mathcal{S}up((\mathtt{X} \cup \mathtt{Y})_{\mathcal{C}}) \geq \epsilon$ and $\mathcal{C}onf(\mathtt{X}_{\mathcal{C}} \longrightarrow \mathtt{Y}_{\mathcal{C}}) \geq \delta$.

Assignment:

Implement an algorithm that solves the CARM problem, i.e., given support/confidence thresholds $0 \leq \epsilon, \delta \leq 1$ and a dataset D (inputs) enumerates a complete set of Clustering Association Rules on D which holds with support at least ϵ and confidence at least δ. Test the resulting implementation by extracting the rules on a dataset of your choice such as AirQuality on the UCI machine learning dataset repository.