Contribution of low support association rules in understanding the mined knowledge

Iztok Fister, Andres Iglesias, Akemi Galvez, Vili Podgorelec, Iztok Fister Jr.

University of Maribor

ACDSA, February 2024

Presentation agenda

Motivation

- Paper aims to explore the impact of low-support association rules on understanding the knowledge domain through analysis of mined association rule archives.
- Simple rules have fewer attributes, while complex ones involve almost all. Fitness value, a linear combination of support, confidence, and inclusion, converges to one with increasing generation numbers.
- Evolution process may replace some attributes in rules, contributing more to fitness values. Numerical attributes can cover the entire feasible value domain with proposed intervals.

NARM Problem Definition

- The NARM problem is mathematically defined as follows:
- Let $T = \{t_1, \ldots, t_N\}$ be a set of transactions.
- Each transaction contains a subset of features (itemset) $F = \{A_1, \ldots, A_M\}$.
- Features can be discrete or numerical (integer or real).
- Discrete features: $A^{(dis)} = \{a_1, \ldots, a_Q\}$.
- Numerical features: $A^{(num)} \in [lb, ub]$, where lb and ub are lower and upper bounds.

Association Rule Definition

- Association rule as an implication:

$$X \Rightarrow Y, \tag{1}$$

- X and Y are two itemsets.
- It holds that $X \cap Y = \emptyset$ (no common elements).
- Variables:
  - M: Number of attributes.
  - N: Number of transactions in the database.
  - Q: Number of attributes in the set $A^{(dis)}$.

Interestingness Measures

- Several interestingness measures are defined for identifying and evaluating association rules.
- Commonly used measures include support and confidence:

$$\mathrm{supp}(X \Rightarrow Y) = \frac{|t_i|t_i \in X \wedge t_i \in Y|}{N}, \tag{2}$$

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{Support}(X \cup Y)}{\mathrm{Support}(X)}, \tag{3}$$

- $\mathrm{supp}(X \Rightarrow Y) \geq S_{\mathrm{min}}$ denotes minimum support.
- $\mathrm{conf}(X \Rightarrow Y) \geq C_{\mathrm{min}}$ denotes minimum confidence.
- Only rules with support and confidence higher than S_{min} and C_{min} are considered.

Inclusion Interestingness Measure

- Additionally, an inclusion $\mathrm{incl}(X \Rightarrow Y)$ NARM interestingness measure is defined:

$$\mathrm{incl}(X \Rightarrow Y) = \frac{\mathrm{ante}(X \Rightarrow Y) + \mathrm{cons}(X \Rightarrow Y)}{M}, \qquad (4)$$

- $\mathrm{ante}(X \Rightarrow Y)$: Set of objects in the antecedent.
- $\mathrm{cons}(X \Rightarrow Y)$: Set of objects in the consequent.
- Mathematically expressed functions:

$$\mathrm{ante}(X \Rightarrow Y) = \{o_{\pi_j} | \pi_j < \mathrm{Cp}_i^{(t)} \wedge \mathrm{Th}(\mathrm{Attr}_{\pi_j}^{(t)}) = \mathrm{enabled}\},$$
$$\mathrm{cons}(X \Rightarrow Y) = \{o_{\pi_j} | \pi_j \geq \mathrm{Cp}_i^{(t)} \wedge \mathrm{Th}(\mathrm{Attr}_{\pi_j}^{(t)}) = \mathrm{enabled}\}.$$

- $\mathrm{incl}(X \Rightarrow Y)$ estimates how many features contribute to the association rule among all.

Analysing an archive of mined association rules

An archive of mined association rules is analyzed in this section. The aim of this analysis was three-fold:

- To identify the distribution of attributes within the UCI ML dataset in order to determine the complexity of the problem.
- To detect a phenomenon of covering the whole interval of possible values by NARM solver.
- To indicate the problem of the features being disappeared in the association rules.

All the analyses were performed on the Abalone dataset taken from the UCI ML repository

Impact of Distributions on NARM Metrics

- Abalone dataset presented as random variables with $Q = 10$ sample points.
- Frequencies of random variables not normally distributed.
- Different impact on NARM metrics calculation:
  - Discrete features limited by attribute set size.
  - Numeric features limited by interval $[lb, ub]$ proposed by NARM solver.
- Number of sample points Q determined based on numeric and discrete features.
- Evolutionary search process performance influenced by dataset composition.
- More numeric features ease mining better association rules due to NARM solver flexibility.
- More discrete values make the search for better rules complex in evolutionary search.

Detecting Domain Coverage by Numerical Features

- uARMSolver adapts proposed intervals of numerical features towards the whole domain of feasible values.
- Expectation that proposed intervals will match the entire domain with evolutionary search maturity.
- Analysis focuses on two association rules: AR-1 and AR-2.
- The covering of the whole domain increases the support metric to one.

Detecting Disappeared Features in Association Rules

- **Phenomenon**: Specific features disappear from association rules due to small support or confidence.
- **Goal**: Analyze where and why features disappear, and understand their contribution to hidden knowledge in the transaction database.
- Typically, disappearing features are discrete, limited by the number of different classes.
- Numerical features have no such limitation; their support metrics could converge to one by widening intervals.

Example of Disappeared Feature in Association Rule AR-3

- **Association Rule AR-3 (28.59% Coverage):**
  - Antecedent: Six attributes including 'Sex' 'I'.
  - Consequent: Two attributes including 'Rings' [12,21] and 'Shucked weight' [1.2407,1.4480].
  - Total coverage: 29.60%
  - 'Sex' 'I' cannot improve support metric and will be replaced in subsequent generations.
  - AR-3 provides insights into the knowledge about Abalone domain related to the disappearing discrete attribute 'Sex_I'.
- Detecting disappeared features involves sorting the archive by fitness values and identifying association rules with disappearing features.

Conclusion

- Nature-inspired NARM solvers (e.g., uARMSolver) generate a diverse archive of association rules based on fitness function values.
- Study Purpose: Explore additional knowledge from lower-quality association rules.
- Analyses on Abalone UCI ML dataset revealed:
  - NARM metrics (support, confidence) dependent on attribute distribution.
  - Phenomenon of covering the entire domain of feasible values causing feature disappearance.
  - Lower support association rules contribute to understanding mined knowledge, especially those with disappeared features.
- Developed algorithm for detecting interesting association rules of lower quality.
- Implications for NARM Solvers:
  - Limit intervals of numeric attributes considering probability distribution.
  - Future Direction: Explore probability distributions as potential research area.

Questions