Frequent Relational Pattern Mining

Martin Svatoš

October 13, 2016

Motivation for frequent pattern mining

- find set of items that are used frequently together
- market basket analysis
- fraud detection, technical dependence analysis, building classifiers
- association rules

Definitions

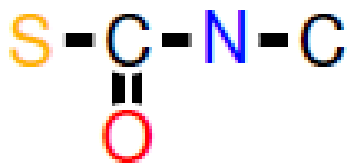
- item base $B = \{i_1, \ldots, i_m\}$
- any subset of B is called an item set, $I \subseteq B$
- transaction database T is a vector of transactions over B (item sets)
- transaction t *covers* item set i iff $i \subseteq t$.
- support of an item set X: $sup(X) = \#$ transactions of T covering X
- association rule (AR) is $X \implies Y$ where X and Y are item sets and $X \cap Y = \emptyset$
- confidence of AR $X \implies Y$: $conf(X \implies Y) = \frac{sup(X \cup Y)}{sup(X)}$
- AR $X \implies Y$ informally: if the transaction covers X, then it is most likely that it also covers Y

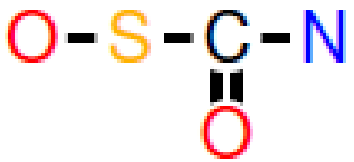
Apriori algorithm [1]

- two phase algorithm for mining frequent patterns
- returns all item sets with support at least sup_{min} (user specified parameter)
- 1) candidate generation
- 2) pruning
- level-wise algorithm by cardinality of item set
- anti-monotone property: no superset of an infrequent item set can be frequent
- another used approach, different from Apriori, is pattern growth

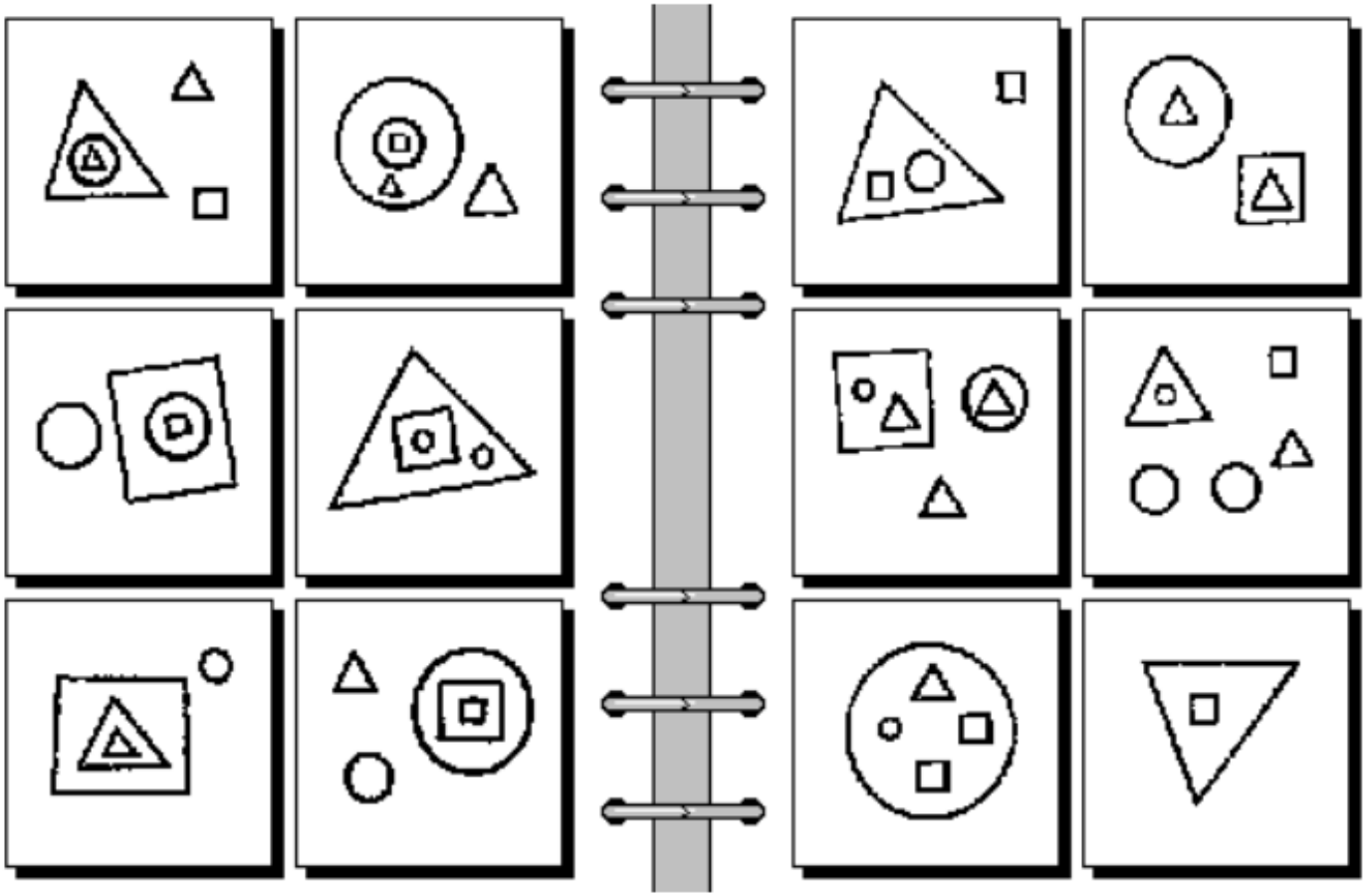
Motivation for relational domain

- mining frequent subgraphs [3]
- use gSpan, Gaston, LEAP, MoFa, Subdue, . . . (pattern growth approach)
- or FSG, FFSM, SPIN, FTOSM, . . . (Apriori approach)





So gSpan, mine frequent patterns from Bongard's problems [2]



- Find out pattern that is common in the left side but is missing in the right side.
- Converting the picture to graph?

Warmr [4]

- ILP method based on Apriori *generate & prune* approach
- finds relational frequent patterns, precisely Datalog queries
- *atom set* instead of item set
- support equals to number of different bindings of the query
- monotone specialization
- candidates are generated by extending older ones by allowed extension (language bias)
- does not follow level-wise generation of candidates
- *occurrence check*

Warmer [6]

- aimed to be more efficient than Warmr
- diagonally contained query within another
- to bee more level-wise
- introduces operations for candidate generation: extension, join, selection, projection
- generate a small superset of all possible candidate queries and remove each query of which a generalization is not known to be frequent

Farmer [7]

- the same goal as Warmr
- use different *occurrence check*
- different structure for storing queries
- faster than Warmr

Intermezzo: solving classification problems with Warmr [5]

1. find n frequent patterns
2. construct a boolean matrix M such that m_{ij} is true iff example i contains pattern j; otherwise false
3. learn decision tree from the matrix

Descriptive language bias

- narrows the space of hypothesis

Beyond

- another ILP methods: SPADA
- application of relational frequent patterns: networking, healthcare, sales domains
- relational frequent patterns in stream data: Star FP Stream [8], SWARM

Q & A

Bibliography I


Rakesh Agrawal, Ramakrishnan Srikant, et al. "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994, pp. 487–499.


M. M. Bongard. *Pattern Recognition*. Rochelle Park, N.J.: Hayden Book Co.,Spartan Books, 1970.


Giovanni Da San Martino and Alessandro Sperduti. "Mining structured data". In: *IEEE Computational Intelligence Magazine* 5.1 (2010), pp. 42–49.


Luc Dehaspe and Luc De Raedt. "Mining association rules in multiple relations". In: *International Conference on Inductive Logic Programming*. Springer. 1997, pp. 125–132.

Bibliography II


Carlos Abreu Ferreira, João Gama, and Vıtor Santos Costa. "Ruse-warmr: Rule selection for classifier induction in multi-relational data-sets". In: *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. Vol. 1. IEEE. 2008, pp. 379–386.


Bart Goethals and Jan Van den Bussche. "Relational association rules: getting warmer". In: *Pattern Detection and Discovery*. Springer, 2002, pp. 125–139.


Siegfried Nijssen and Joost Kok. "Faster association rules for multiple relations". In: *International Joint Conference on Artificial Intelligence*. Vol. 17. 1. Citeseer. 2001, pp. 891–896.


Andreia Silva and Cláudia Antunes. "Multi-relational pattern mining over data streams". In: *Data Mining and Knowledge Discovery* 29.6 (2015), pp. 1783–1814.