Association Rule Mining, Sequential Pattern Mining

Core Methods in Educational Data Mining

Valdemar Švábenský | University of Pennsylvania | Oct 20, 2022

Based on the slides created by Ryan Baker for the EDUC 691 course in Spring 2019

Previous assignment (Basic: BKT)

- Questions? Comments? Concerns?

Today’s topics

- Association rule mining
  - a.k.a. Association rule learning
- Sequential pattern mining
  - a.k.a. Sequence mining
- Two related **data mining techniques**
- Finding frequently occurring patterns in a dataset
  - Describing past data
  - **Not** making predictions about the future

Part 1/2:
Association Rule Mining (ARM)

What is ARM?

- Automated discovery of **if-then patterns** in a dataset
  - X → Y means “if X, then Y”
  - **Is it the same as Y → X?**

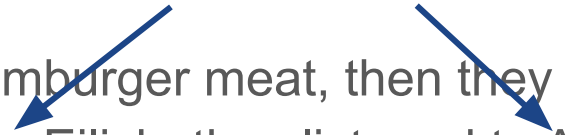
What is ARM?

- Automated discovery of **if-then patterns** in a dataset
  - X → Y means “if X, then Y”
  - **Is it the same as Y → X?**
- Examples:
  - If a grocery shopper bought hamburger meat, then they bought buns
  - If a Spotify user listened to Billie Eilish, they listened to Ariana Grande
  - If a student asked for hints to the first three tasks, and the time spent on each task was more than 1 hour, they did not finish the last task
  - **Can you think of other examples?**

What is ARM?

- Automated discovery of **if-then patterns** in a dataset
  - X → Y means “if X, then Y”
  - **Is it the same as Y → X?**
- Examples:
  - If a grocery shopper bought hamburger meat, then they bought buns
  - If a Spotify user listened to Billie Eilish, they listened to Ariana Grande
  - If a student asked for hints to the first three tasks, and the time spent on each task was more than 1 hour, they did not finish the last task
  - **Can you think of other examples?**

Works with qualitative (non-numerical) data!



What is ARM?

- Automated discovery of **if-then patterns** in a dataset
  - X → Y means “if X, then Y”
  - **Is it the same as Y → X?**
- Examples:
  - If a grocery shopper bought hamburger meat, then they bought buns
  - If a Spotify user listened to Billie Eilish, they listened to Ariana Grande
  - If a student asked for hints to the first three tasks, and the time spent on each task was more than 1 hour, they did not finish the last task
  - **Can you think of other examples?**

Works with qualitative (non-numerical) data!

X (or Y) can be a complex condition

Why is ARM useful?

- Making sense of your data
  - What student actions/events occurred together?
  - Which of those occurred often?
- Generating hypotheses from your data for further research
- Finding actionable insights
  - Providing basis for recommendations
- **Can you think of other use cases?**

Two key metrics of association rules

- What is **support**?
- Why is it useful?
- What values can it have?

- What is **confidence**?
- Why is it useful?
- What values can it have?

- May inflate the importance of the rule if X and Y each have high support alone – the rule may be simply due to chance

Exercise

student_id	took_stats	took_DM
A	1	1
B	0	1
C	1	1
D	1	0
E	0	0
F	0	1
G	0	1
H	1	1

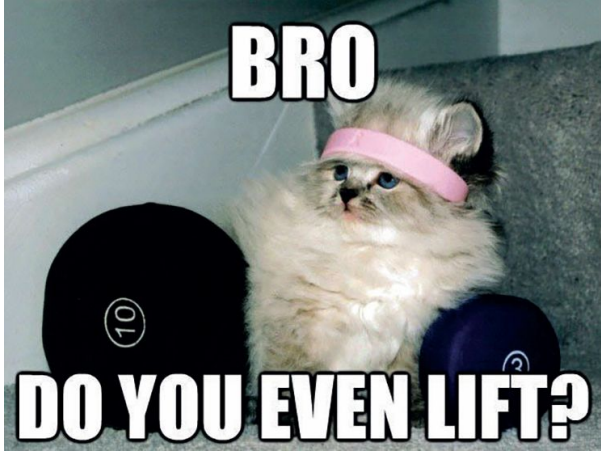
- X = “Student took a stats class” (took_stats = 1)
- Y = “Student took a data mining class” (took_DM = 1)
- Consider a rule X → Y
  - **support(X → Y) = ?**
  - **confidence(X → Y) = ?**

Interestingness metrics for association rules

- **Why do we need** to measure if a rule is “interesting”?
- Do you consider these rules interesting?
  - Students who took a course took its prerequisites (Vialardi et al., 2009)
  - Students who do poorly on the exams fail the course (El-Halees, 2009)
- What are some **metrics for interestingness** of a rule?

Interestingness metrics: Lift

$$\frac{\text{confidence}(X \rightarrow Y)}{\text{support}(Y)}$$



- Expresses the degree to which *Y is more common to appear when X is present* as opposed to *only Y appearing*
  - Lift > 1: X and Y are positively associated
  - Lift = 1: the occurrence of X does not impact whether Y occurs
  - 0 ≤ Lift < 1: X and Y are negatively associated
- **Compute it on the example dataset**

Interestingness metrics: Cosine

$$\frac{\text{support}(X \rightarrow Y)}{\text{sqrt}(\text{support}(X) * \text{support}(Y))}$$

- Ranges from 0 to 1, **why?**
- Expresses co-ocurrence
  - Closer to 1, the more transactions containing X *also contain Y*
  - Closer to 0, the more transactions contain X *without containing Y*
  - Over 0.65 is desirable
- **Compute it on the example dataset**

Interestingness metrics: Jaccard

$$\frac{\text{support}(X \rightarrow Y)}{\text{support}(X) + \text{support}(Y) - \text{support}(X \rightarrow Y)}$$

- Expresses the degree to which having X and Y *together* is more common than having *either X or Y but not both*
  - **What is the possible range of values and why?**
  - Over 0.5 is desirable
- **Compute it on the example dataset**

Algorithms for ARM

- Most straightforward algorithm: **Apriori** (1994)
  - Many others: FP-Growth, MagnumOpus, Closet…
  - Foundations: GUHA (1966), Czech mathematician P. Hájek et al. https://link.springer.com/article/10.1007/BF02345483
- Only rules that satisfy the **user-defined thresholds** *MinSup* and *MinConf*
- Python:
  - https://pypi.org/project/apyori/
  - https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
- R: https://www.datacamp.com/tutorial/market-basket-analysis-r

Making sense of association rules: tabulation

- Example (not actual rules):

Rule	**Sup**	**Conf**
If a student asked for two hints in a row, the time between the clicks in the system was short	0.33	0.42
If the time between two clicks in the system was short, the student asked for two hints	0.21	0.28

Making sense of association rules: visualization

- R package *arules*

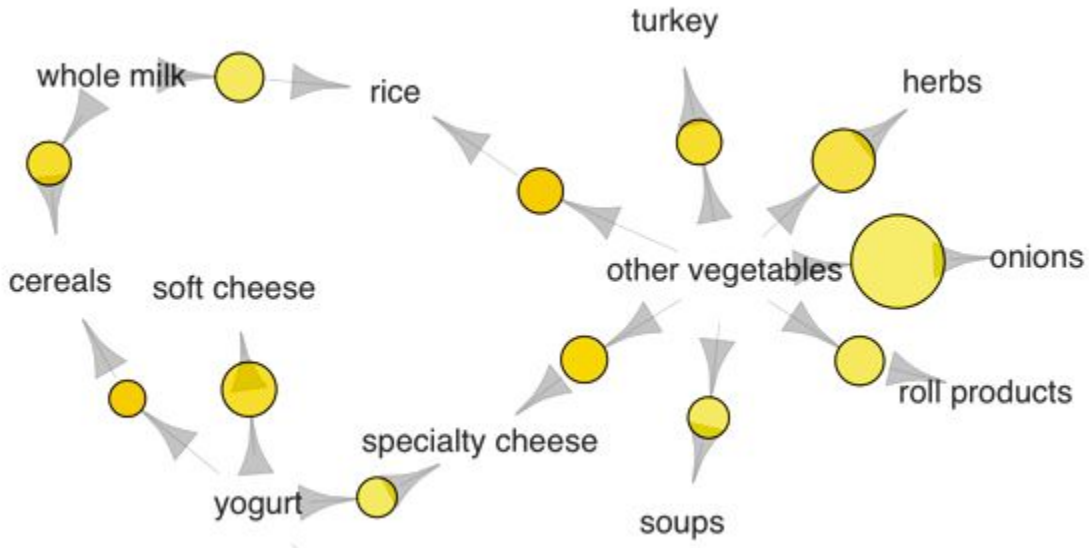


Image credit: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Summary and closing remarks

- Association rule = “if X, then Y” pattern in a dataset
  - Many algorithms exist for finding them automatically, e.g., Apriori
  - Each rule has a certain support and confidence (from 0 to 1)
  - The rules can be tabulated or visualized in a graph
- Interestingness = “does this rule appear relevant?”
  - Helpful metrics: lift, cosine, Jaccard, and many others
  - The researcher must decide if the rule is actually useful to keep
- **Questions?**
- **Comments on the readings?**

Resources, further reading, and code examples

- Papers on the course website
- https://is.muni.cz/auth/th/cxvr2/tkacik-thesis.pdf (definitions)
- https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html/2 (Apriori explained on an example)
- https://medium.com/analytics-vidhya/association-analysis-in-python-2b955d0180c (Python implementation example)
- https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6 (technical analysis)

Part 2/2:
Sequential Pattern Mining (SPM)

What is SPM?

- Automated discovery of **subsequences** in a set of sequences (= discrete, temporal data)
  - $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_N$ is a sequence of N actions
  - The dataset consists of several sequences
  - Which parts of sequences are common?
- **Ordering of items** (actions) in the input data **matters**
  - Unlike for ARM, which looks at each “transaction” as an unordered bag of items

Example

Assume the following encoding of student actions:

- W = **work** on a task
- A = submit an **answer**
- H = reveal a **hint**

Data about three students:

- WAHWA
- HHHA
- WAWAHWAHWA

Again, the data can be qualitative!

Example

Assume the following encoding of student actions:

- W = **work** on a task
- A = submit an **answer**
- H = reveal a **hint**

Data about three students:

- WA**H**W**A**
- HH**HA**
- WAWA**H**W**AH**W**A**

Common subsequence: **HA**

(by default, the sequence can be interleaved by other actions)

Support = 3/3 = 1

(doesn’t matter that HA appeared twice in the last sequence)

Why is SPM useful?

- What are typical consecutive actions of students?
  - Do they typically use the learning system in a certain way?
  - Do they get stuck repeating the same actions?
- Generating hypotheses from your data for further research
- **Can you think of other use cases?**

Exercise

In the following dataset consisting of 3 items, what is the **support** of the subsequence (a) ABC, (b) BD?

- AABABABAD
- CCABACABC
- BDDBDDBCD

Algorithms for SPM

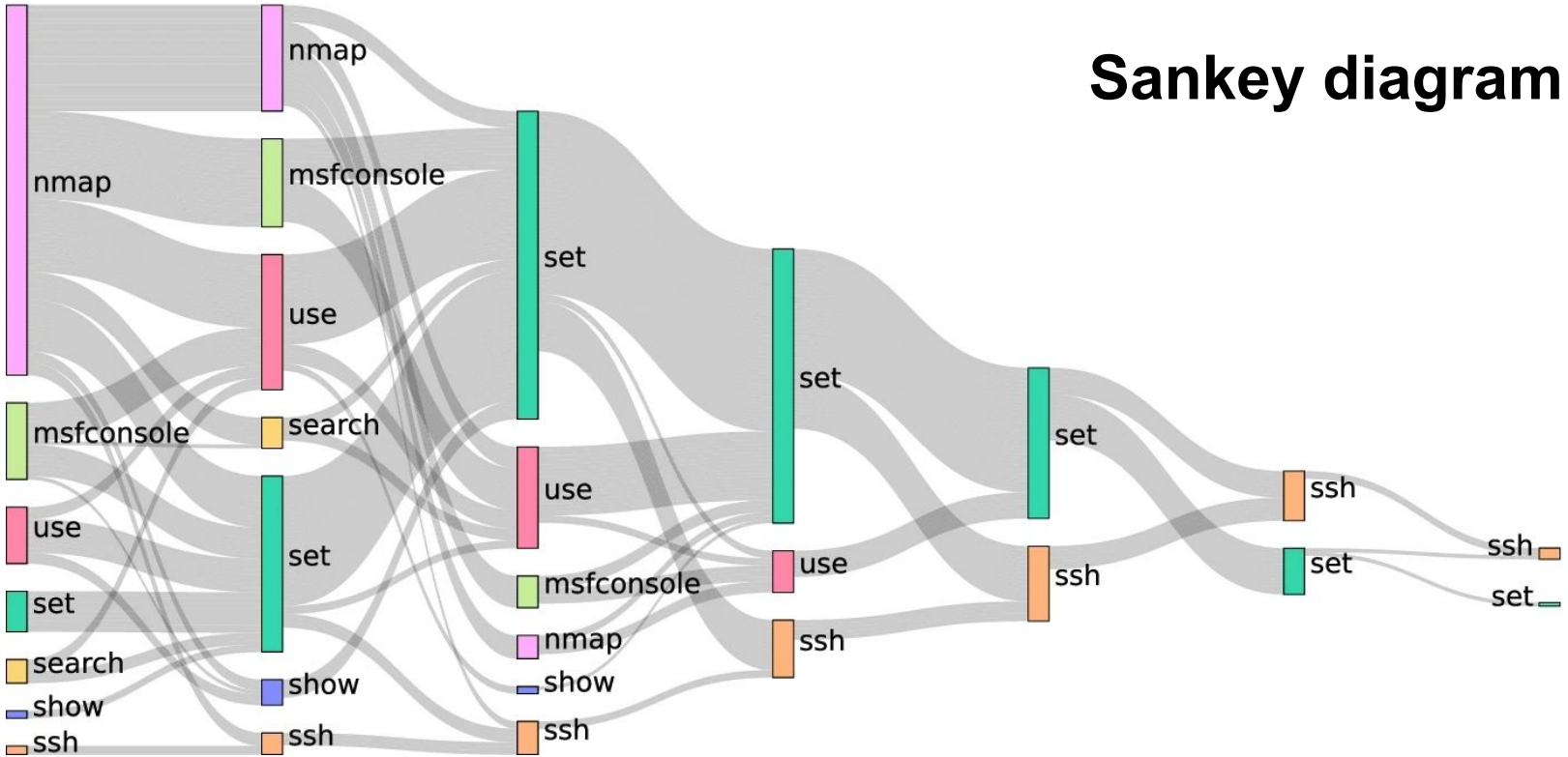
- There are many of them, but they produce the same result
  - GSP, Spade, PrefixSpan, CloFast, Clasp…
- Only rules that satisfy the **user-defined threshold** *MinSup*
- Java:
  - https://philippe-fournier-viger.com/spmf/index.php?link=download.php (can be called inside a Python script)

Making sense of sequential patterns: tabulation

- Example:

Rule	**Sup**
Open Python console → Install modules → Investigate the error that occurred → Ask a question on Piazza	0.10
Open RapidMiner → Preprocess data → Train model → Cross-validate → Submit solution on Piazza	0.15

Making sense of sequential patterns: visualization



Sankey diagram

Summary and closing remarks

- Sequential pattern = frequently occurring sequence of items (e.g., actions, events) in your dataset
  - Many algorithms for finding them automatically, e.g., CloFast
  - Each pattern has a certain support (from 0 to 1)
  - The patterns can be tabulated or visualized in a graph
  - Sequence = order matters
- **Questions?**
- **Comments on the readings?**
- **More ideas for the educ. applications of ARM/SPM?**

Quiz time!

- On your phone, go to **play.blooket.com**
- Enter the ID code shown on the projector
- Choose your nickname (SFW please) and avatar
- Answer multiple-choice questions: both accuracy and speed count

Image credit: https://www.blooket.com/