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
  - $X \rightarrow Y$ means “if $X$, then $Y$”
  - Is it the same as $Y \rightarrow X$?
What is ARM?

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

<table>
<thead>
<tr>
<th>student_id</th>
<th>took_stats</th>
<th>took_DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **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

\[
\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cap Y)}{\text{support}(Y)}
\]

- 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
  - \( 0 \leq \text{Lift} < 1 \): \( X \) and \( Y \) are negatively associated

- **Compute it on the example dataset**
Interestingness metrics: Cosine

\[
\text{support}(X \rightarrow Y) = \sqrt{\text{support}(X) \times \text{support}(Y)}
\]

- Ranges from 0 to 1, why?
- Expresses co-occurrence
  - 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

\[
\text{support}(X \rightarrow Y) = \frac{\text{support}(X \rightarrow Y)}{\text{support}(X) + \text{support}(Y) - \text{support}(X \rightarrow Y)}
\]

- 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.

- Only rules that satisfy the **user-defined thresholds** `MinSup` and `MinConf`

- Python:
  - [https://pypi.org/project/apyori/](https://pypi.org/project/apyori/)
  - [https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/](https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/)

- R: [https://www.datacamp.com/tutorial/market-basket-analysis-r](https://www.datacamp.com/tutorial/market-basket-analysis-r)
Making sense of association rules: tabulation

- Example (not actual rules):

<table>
<thead>
<tr>
<th>Rule</th>
<th>Sup</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>If a student asked for two hints in a row, the time between the clicks in the system was short</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td>If the time between two clicks in the system was short, the student asked for two hints</td>
<td>0.21</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Making sense of association rules: visualization

- R package *arules*

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://medium.com/analytics-vidhya/association-analysis-in-python-2b955d0180c](https://medium.com/analytics-vidhya/association-analysis-in-python-2b955d0180c) (Python implementation example)
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
- HHHHA
- WAWAHWAHWA

Again, the data can be qualitative!
Example

Assume the following encoding of student actions:

- \( W = \text{work} \) on a task
- \( A = \text{submit an answer} \)
- \( H = \text{reveal a hint} \)

Data about three students:

- WA\text{HWA}
- HHHA
- WA\text{WAHWA}

Common subsequence: \( \text{HA} \)
(by default, the sequence can be interleaved by other actions)

Support = \( \frac{3}{3} = 1 \)
(doesn’t matter that \( \text{HA} \) 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
- BDDBDDBCDDB
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:
Making sense of sequential patterns: tabulation

**Example:**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Sup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Python console → Install modules → Investigate the error that occurred → Ask a question on Piazza</td>
<td>0.10</td>
</tr>
<tr>
<td>Open RapidMiner → Preprocess data → Train model → Cross-validate → Submit solution on Piazza</td>
<td>0.15</td>
</tr>
</tbody>
</table>
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/