InterVisAR: An Interactive Visualization for Association Rule Search

Chih-Wen Cheng, Ying Sha, May D. Wang

Georgia Institute of Technology
Oct. 5th, 2016
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
Association Rule Mining – Concept

A popular method for discovering interesting relations among any variables in large databases [1]
– \{Butter, bread\} \Rightarrow \{Milk\}

Evidence-based: \(X\) (antecedent) \(\Rightarrow\) \(Y\) (consequent)[1]
– We have seen \(X\) (e.g., observations), ARM quantitatively tells us its implication (i.e. possibility) of \(Y\) (e.g., targeted outcomes)

Example healthcare data

<table>
<thead>
<tr>
<th>PID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Drug_\text{A}, Cond_\text{B}</td>
</tr>
<tr>
<td>2</td>
<td>Drug_\text{A}, Drug_\text{B}, Cond_\text{A}, Cond_\text{C}</td>
</tr>
<tr>
<td>3</td>
<td>Cond_\text{B}, Drug_\text{B}, Cond_\text{A}, Cond_\text{D}</td>
</tr>
<tr>
<td>4</td>
<td>Drug_\text{A}, Cond_\text{B}, Drug_\text{B}, Cond_\text{A}</td>
</tr>
<tr>
<td>5</td>
<td>Drug_\text{A}, Cond_\text{B}, Drug_\text{B}, Cond_\text{D}</td>
</tr>
</tbody>
</table>

Rule: \(\{\text{Drug}_\text{A}\} \Rightarrow \{\text{Cond}_\text{A}\}\)

\[\text{Supp} = \frac{2}{5} = 40\%\]

\[\text{Conf} = \frac{2}{4} = 50\%\]

1. Agrawal, et al., ICDE, 1995
Tabulated Association Rules

$N$ items can generate up to $2^N - 1$ rules...

<table>
<thead>
<tr>
<th>Antecedents</th>
<th>Supp</th>
<th>Cont</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPWORTH_SCORE &gt;= 10.3 &amp; FAD_TOTALCOMM &gt;= 2.3 &amp; GAD7_SCORE &gt;= 6.1</td>
<td>1%</td>
<td>43%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &gt;= 10.3 &amp; FAD_TOTALCOMM &gt;= 2.3 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>1%</td>
<td>27%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; ESSI_SCORE &lt;= 23.9 &amp; PSSE_SCORE &gt;= 25.4</td>
<td>3%</td>
<td>38%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; ESSI_SCORE &lt;= 23.9 &amp; GAD7_SCORE &gt;= 6.1</td>
<td>4%</td>
<td>60%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; ESSI_SCORE &lt;= 23.9 &amp; SPEXS_SCORE &lt;= 60.8</td>
<td>2%</td>
<td>27%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; EPWORTH_SCORE &lt;= 10.3 &amp; ESSI_SCORE &lt;= 23.9</td>
<td>1%</td>
<td>27%</td>
</tr>
<tr>
<td>ESSI_SCORE &lt;= 23.9 &amp; GAD7_SCORE &gt;= 6.1 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>3%</td>
<td>47%</td>
</tr>
<tr>
<td>ESSI_SCORE &lt;= 23.9 &amp; PSSE_SCORE &lt;= 25.4 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>3%</td>
<td>52%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &lt;= 10.3 &amp; ESSI_SCORE &lt;= 23.9 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>1%</td>
<td>25%</td>
</tr>
<tr>
<td>ESSI_SCORE &lt;= 23.9 &amp; GAD7_SCORE &gt;= 6.1 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>3%</td>
<td>83%</td>
</tr>
<tr>
<td>ESSI_SCORE &lt;= 23.9 &amp; GAD7_SCORE &gt;= 6.1 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>1%</td>
<td>77%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &lt;= 10.3 &amp; ESSI_SCORE &lt;= 23.9 &amp; GAD7_SCORE &gt;= 6.1</td>
<td>0%</td>
<td>18%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &lt;= 10.3 &amp; ESSI_SCORE &lt;= 23.9 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>3%</td>
<td>41%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; GAD7_SCORE &gt;= 6.1 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>3%</td>
<td>49%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; PSSE_SCORE &lt;= 25.4 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>1%</td>
<td>26%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; EPWORTH_SCORE &lt;= 10.3 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>1%</td>
<td>26%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; GAD7_SCORE &gt;= 6.1 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>3%</td>
<td>56%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; EPWORTH_SCORE &lt;= 10.3 &amp; GAD7_SCORE &gt;= 6.1</td>
<td>2%</td>
<td>58%</td>
</tr>
<tr>
<td>BDI_SCORE &gt;= 9.1 &amp; EPWORTH_SCORE &lt;= 10.3 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>1%</td>
<td>26%</td>
</tr>
<tr>
<td>GAD7_SCORE &gt;= 6.1 &amp; PSSE_SCORE &lt;= 25.4 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>3%</td>
<td>45%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &lt;= 10.3 &amp; GAD7_SCORE &gt;= 6.1 &amp; PSSE_SCORE &lt;= 25.4</td>
<td>1%</td>
<td>21%</td>
</tr>
<tr>
<td>EPWORTH_SCORE &lt;= 10.3 &amp; PSSE_SCORE &lt;= 25.4 &amp; SPEXS_SCORE &lt;= 60.6</td>
<td>1%</td>
<td>84%</td>
</tr>
</tbody>
</table>
Conventional AR Visualization Techniques

Scatter Plot
- Rule with extreme values
- Each dot is a rule

Matrix-based Plot

Graph-based Plot
- Soda
- Soda + Bread
- Ham & Bread
- Bread
- Ham
- Ham & Soda

Matrix with 371 rules

Antecedent (lhs) vs. Consequent (lhs)
Challenges of Current AR Visualization

- Designed to find extreme cases or summarize the distribution of all rules
  - None of them is designed for rule searching
- Require mapping process => Not intuitive
- Reasoning process => Tweak rules to find new rules
- Non-factorial property
  - The demand of rule search mainly depends on the size of the extracted rule pool that may contain up to $C_i^N$ rules.
  \[
  C_4^8 = 70 \text{ v.s. } C_4^9 = 126 \quad C_4^8 = 70 \text{ v.s. } C_7^8 = 8
  \]
  - The demand of using a rule searching technique should only monotonically increases for longer antecedents, instead of being factorially influenced by the number of possible items $N$ (i.e., $C_i^N$).
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
InterVisAR - Visualization

Potential Items

- High BP
- Low SpO2
- Low HR
- Low Temp.
- Low Plateletes
- Low Lactic Acid
- High HR

(a) Percentage (%)

Potential Items

- High BP
- Low Temp.
- Low Lactic Acid
- Low HR
- High HR
- Low Plateletes

(b) Percentage (%)

Selected Items

- Low SpO2

Reset
InterVisAR - Algorithm

Algorithm InterARMVis(\( \mathcal{R} \), \( Supp_{\text{min}} \), \( Conf_{\text{min}} \), sort_type)

Input: \( \mathcal{R} \) is the extracted rule set, \( Supp_{\text{min}} \) is the minimum support, \( Conf_{\text{min}} \) is the minimum confidence, and sort_type is the type of sorting

1. draw \( Supp_{\text{min}} \) and \( Conf_{\text{min}} \);
2. sort \( \mathcal{R} \) by sort_type;
3. set \( r_{\text{current}} = \emptyset \);
4. Update \( r_{\text{current}} \cdot \text{conf} \) and \( r_{\text{current}} \cdot \text{supp} \);
5. \( W = \text{extractRulesOfLength}(r_{\text{current}}, \mathcal{R}) \);
6. while \( W \neq \emptyset \) do
7.   clear the current plot;
8.   for each rule \( r \) in \( W \) do
9.     Add the item \( \{r.a - r_{\text{current}}.a\} \) to the row label;
10.    Add a support bar from 0\% to \( r.\text{supp} \);
11.   \( r_{\text{max}} = \text{extractMaxOfItem}(r.a, \mathcal{R}) \);
12.   Add a confidence line from \( r.\text{conf} \) to \( r_{\text{max}}.\text{conf} \);
13.   \text{significant_change} = \text{BinomTest}(r_{\text{current}}, r);
14.   if \text{significant_change} == Yes then
15.     Plot solid circle on the \( r.\text{conf} \);
16.   else Plot hollow circle on the \( r.\text{conf} \);
17.   while an item \( i \) is selected from \( r_{\text{current}}.a \) do
18.     \( r_{\text{current}}.a = \{r_{\text{current}}.a - i\}; \) goto line 4;
19.   while an rule \( r \) is selected do
20.     \( r_{\text{current}} = r; \) goto line 4;
21.   while reset goto line 3;
22.   while reset goto line 3;
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules
  (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
Predictive Health

“A transformation towards maintaining health (rather then treating diseases) by proactively predicting health-related events and disease development, and providing early and persistent interventions before being clinically overt.” [1]

Center for Health Discovery and Well Being (CHDWB®)

- 2,637 de-identified health reports from 696 healthy participants with 906 health measurements
  - Questionnaires
  - Psychological scales
  - Physical measurements
  - Lab tests

1. Brigham K, et al., 2012
Using Visualization To Improve the Effectiveness of Rule Finding

• Rule Search
  – Finding a rule with a given set of antecedent items.

\[ I_1, I_2, I_3 \Rightarrow O \]

• Next-level Item Search
  – Given a set of antecedent items, find the next item that produces highest confident.

\[ I_1, I_2, I_3 \Rightarrow O \]

\[ I_1, I_2 + I_4 \Rightarrow O \]

\[ I_1, I_2 + I_3 + I_4 + I_5 \Rightarrow O \]
### Use Case 1: Prevent

<table>
<thead>
<tr>
<th>Existing 1st Disorder</th>
<th>Potential 2nd Disorder</th>
<th>Change of %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression (26%)</td>
<td>Self-deficiency</td>
<td>+26%</td>
</tr>
<tr>
<td></td>
<td>Dysfunctional family</td>
<td>+23%</td>
</tr>
<tr>
<td></td>
<td>Lack of social support</td>
<td>+9%</td>
</tr>
<tr>
<td></td>
<td>Anxiety</td>
<td>+3%</td>
</tr>
</tbody>
</table>

Provide proactive advice to prevent disorders of self-deficiency and family dysfunction because they increase the % most significantly.

### Use Case 2: Prioritize

<table>
<thead>
<tr>
<th>Existing Disorders</th>
<th>Remove Disorder</th>
<th>Change of %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-deficiency Depression</td>
<td>Lack of social support</td>
<td>-37%</td>
</tr>
<tr>
<td>Lack of social support</td>
<td>Self-deficiency</td>
<td>-20%</td>
</tr>
<tr>
<td>Anxiety (99%)</td>
<td>Depression</td>
<td>-17%</td>
</tr>
<tr>
<td></td>
<td>Anxiety</td>
<td>-16%</td>
</tr>
</tbody>
</table>

Focus and set action plans for lack of social support because it can significantly drop the risk by 37%.

Dataset - Seven Possible Antecedent Items

<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDI_SCORE&gt;=9.1</td>
</tr>
<tr>
<td>ESSI_SCORE&lt;=23.9</td>
</tr>
<tr>
<td>EPWORTH_SCORE&gt;=10</td>
</tr>
<tr>
<td>SPEXS_SCORE&lt;=60.6</td>
</tr>
<tr>
<td>FAD_TOTALCOMM&gt;=2.3</td>
</tr>
<tr>
<td>GAD7_SCORE&gt;=6.1</td>
</tr>
<tr>
<td>PSSE_SCORE&gt;=25.4</td>
</tr>
</tbody>
</table>

Score Ranges of 7 Psychological Assessments

Use different combinations of antecedent items to predict the possibility of mental illness

<table>
<thead>
<tr>
<th>Length</th>
<th>Example Antecedent</th>
<th>No. of Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GAD7&gt;=6.1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>ESSI&lt;=23.9 &amp; PSSE&gt;=25.4</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>BDI&gt;=9.1 &amp; FAD&gt;=2.3 &amp; GAD7&gt;=6.1</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>BDI&gt;=9.1 &amp; EPWORTH&gt;=10.3 &amp; FAD&gt;=2.3 &amp; GAD7&gt;=6.1</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>EPWORTH&gt;=10.3 &amp; FACT&lt;=60.6 &amp; FAD&gt;=2.3 &amp; GAD7&gt;=6.1 &amp; PSSE&gt;=25.4</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>BDI&gt;=9.1 &amp; EPWORTH&gt;=10.3 &amp; ESSI&lt;=23.9 &amp; FAD&gt;=2.3 &amp; GAD7&gt;=6.1 &amp; PSSE&gt;=25.4</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>BDI&gt;=9.1 &amp; EPWORTH&gt;=10.3 &amp; ESSI&lt;=23.9 &amp; FACT&lt;60.6 &amp; FAD&gt;=2.3 &amp; GAD7&gt;=6.1 &amp; PSSE&gt;=25.4</td>
<td>1</td>
</tr>
</tbody>
</table>

Total 127
Six Rule Presentations

- 24 subjects for user evaluation study

<table>
<thead>
<tr>
<th>Table-based Rules</th>
<th>Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T</strong></td>
<td><em>VisC</em></td>
</tr>
<tr>
<td>A table with rules ascendingly sorted by antecedent length.</td>
<td>InterVisAR with descending sorting by confidence values of next-level items.</td>
</tr>
<tr>
<td><strong>TO</strong></td>
<td><em>VisA</em></td>
</tr>
<tr>
<td>A table with rules ascendingly sorted by antecedent length. Rules with the same length were further sorted alphabetically by itemset.</td>
<td>InterVisAR with alphabetical (a to Z) sorting by names of next-level items.</td>
</tr>
<tr>
<td><strong>TL</strong></td>
<td></td>
</tr>
<tr>
<td>A table with a slider bar controlling the length of displayed antecedents. Rules were descendingly sorted by confidence.</td>
<td></td>
</tr>
<tr>
<td><strong>TLO</strong></td>
<td></td>
</tr>
<tr>
<td>A table with a slider bar controlling the length of displayed antecedents. Rules were alphabetically sorted by itemsets.</td>
<td></td>
</tr>
</tbody>
</table>

- Queries with rules with different lengths
- Record *completion time* and *accuracy*
- Accurate Rules Per Minute (ARPM)

\[
ARPM = \frac{60}{\text{completion time (in sec)}} \times \text{accuracy}
\]
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
User Study – Completion Time

Rule Search

I_1, I_2, I_3 ⇒ O

Next-Item Search

I_1, I_2 + I_3 ⇒ O
+ I_4
+ I_5
User Study - Accuracy

Rule Search

\[ I_1, I_2, I_3 \Rightarrow O \]

Next-Item Search

\[ I_1, I_2 + I_4 \Rightarrow O \]

\[ + I_3 \]

\[ + I_5 \]

Methods

- Table-based
- Visualization

Rule Length

Table-based Visualization

Rule Length

Table-based Visualization

Methods

- Table-based
- Visualization

Table-based Visualization

Methods

- Table-based
- Visualization

Table-based Visualization

Methods

- Table-based
- Visualization
InterVisAR outperforms all table-based presentations in ARPM.
Outline

• Conventional association rule visualization methods
  – Challenges
• Interactive visualization for association rules (InterVisAR)
• User evaluation experiment
• Results
• Conclusion
Conclusion

- Introduced an interactive association rule visualization
  - Interactively receives user inputs step-by-step until all items in the target rule have been found
  - Delivers most association rule information in a small view without overwhelming manual panning and scrolling

- Not only outperformed table-based rule presentations in terms of efficiency, accuracy, and accurate rule per minute (ARPM) but also satisfied the non-factorial property.

- Participants also expressed strong preferences for InterVisAR because it promises a confortable and pleasing rule-searching solution.
Future Work

• To demonstrate its usability on small screens such as on smartphones
  – Minimum restriction on display size

• To evaluate interVisAR in different real-world dataset so as to validate its viability in comprehensive data-mining settings