Data Mining: Concepts and Techniques

*Mining Frequent Patterns, Associations, and Correlations*

— Chapter 5.4 & 5.6 —

---

**Chapter 5: Mining Frequent Patterns, Association and Correlations**

- Basic concepts and a road map
- Efficient and scalable frequent itemset mining methods
- Mining various kinds of association rules
- From association mining to correlation analysis
- Constraint-based association mining
- Summary

---
Interestingness Measure: Correlations (Lift)

- *play basketball ⇒ eat cereal* [40%, 66.7%] is misleading
  - The overall % of students eating cereal is 75% > 66.7%.
- *play basketball ⇒ not eat cereal* [20%, 33.3%] is more accurate, although with lower support and confidence
- Measure of dependent/correlated events: *lift*

\[
lift = \frac{P(A \cup B)}{P(A)P(B)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Basketball</th>
<th>Not basketball</th>
<th>Sum (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cereal</td>
<td>2000</td>
<td>1750</td>
<td>3750</td>
</tr>
<tr>
<td>Not cereal</td>
<td>1000</td>
<td>250</td>
<td>1250</td>
</tr>
<tr>
<td>Sum(col.)</td>
<td>3000</td>
<td>2000</td>
<td>5000</td>
</tr>
</tbody>
</table>

\[
lift(B, C) = \frac{2000/5000}{3000/5000 \times 3750/5000} = 0.89
\]

\[
lift(B, \neg C) = \frac{1000/5000}{3000/5000 \times 1250/5000} = 1.33
\]

Are *lift* and $\chi^2$ Good Measures of Correlation?

- "*Buy walnuts ⇒ buy milk* [1%, 80%]" is misleading
  - if 85% of customers buy milk
- Support and confidence are not good to represent correlations
- So many interestingness measures? (Tan, Kumar, Sritastava @KDD’02)

\[
lift = \frac{P(A \cup B)}{P(A)P(B)}
\]

\[
all\_conf = \frac{\sup(X)}{\max\_item\_\sup(X)}
\]

\[
\chi^2 = \sum(P(A_i) - E_i) / E_i
\]

\[
coh = \frac{\sup(X)}{|universe(X)|}
\]

<table>
<thead>
<tr>
<th></th>
<th>Milk</th>
<th>No Milk</th>
<th>Sum (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>m, c</td>
<td>~m, c</td>
<td>c</td>
</tr>
<tr>
<td>No Coffee</td>
<td>m, ~c</td>
<td>~m, ~c</td>
<td>~c</td>
</tr>
<tr>
<td>Sum(col.)</td>
<td>m</td>
<td>~m</td>
<td>\Sigma</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DB</th>
<th>m, c</th>
<th>~m, c</th>
<th>m~c</th>
<th>~m~c</th>
<th>lift</th>
<th>all-conf</th>
<th>coh</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1000</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>10,000</td>
<td>9.26</td>
<td>0.91</td>
<td>0.83</td>
<td>9055</td>
</tr>
<tr>
<td>A2</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>100,000</td>
<td>8.44</td>
<td>0.09</td>
<td>0.05</td>
<td>670</td>
</tr>
<tr>
<td>A3</td>
<td>1000</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
<td>100,000</td>
<td>9.18</td>
<td>0.09</td>
<td>0.09</td>
<td>8172</td>
</tr>
<tr>
<td>A4</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1</td>
<td>0.5</td>
<td>0.33</td>
<td>0</td>
</tr>
</tbody>
</table>
Which Measures Should Be Used?

- **lift** and \( \chi^2 \) are not good measures for correlations in large transactional DBs
- **all-conf** or **coherence** could be good measures (Omiecinski, TKDE’03)
- Both **all-conf** and **coherence** have the downward closure property
- Efficient algorithms can be derived for mining (Lee et al., ICDM’03sub)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Measure</th>
<th>Range</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta )</td>
<td>( \delta )-coefficient</td>
<td>-1 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A)P(B)}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
<tr>
<td>( Q )</td>
<td>Yale’s ( Q )</td>
<td>-1 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{P(A)P(B)} )</td>
</tr>
<tr>
<td>( Y )</td>
<td>Yale’s ( Y )</td>
<td>-1 \ldots 1</td>
<td>( \sqrt{\frac{P(A \cup B) - P(A)P(B)}{P(A)P(B)}} )</td>
</tr>
<tr>
<td>( k )</td>
<td>Cohen’s ( k )</td>
<td>-1 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A)P(B)}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
<tr>
<td>( PS )</td>
<td>Piatetsky-Shapiro’s ( PS )</td>
<td>0.25 \ldots 0.75</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{P(A) + P(B) - 2P(A)P(B)} )</td>
</tr>
<tr>
<td>( F )</td>
<td>Certainty factor</td>
<td>-1 \ldots 1</td>
<td>( \max \left( \frac{P(A \cup B) - P(A)P(B)}{P(A)P(B)}, \frac{P(A \cup B) - P(A)P(B)}{P(A)P(B)} \right) )</td>
</tr>
<tr>
<td>( AV )</td>
<td>added value</td>
<td>-0.5 \ldots 1</td>
<td>( \frac{\min(P(A \cup B) - P(A) - P(B) + 1, 0)}{P(A) + P(B) - 2P(A)P(B)} )</td>
</tr>
<tr>
<td>( K )</td>
<td>Kline’s ( K )</td>
<td>-0.33 \ldots 0.38</td>
<td>( \sqrt{\frac{P(A \cup B) - P(A) - P(B) + 1}{P(A)P(B)}} )</td>
</tr>
<tr>
<td>( s )</td>
<td>Goodman-Kruskal’s ( s )</td>
<td>0 \ldots 1</td>
<td>( \frac{\min(P(A \cup B) - P(A)P(B), 0)}{\min(P(A \cup B) - P(A)P(B), 0)} )</td>
</tr>
<tr>
<td>( M )</td>
<td>Mutual Information</td>
<td>0 \ldots 1</td>
<td>( \sum_{x} P(x) \log \frac{P(x \mid y)}{P(x)} + \sum_{y} P(y) \log \frac{P(x \mid y)}{P(x)} )</td>
</tr>
<tr>
<td>( G )</td>
<td>Gini index</td>
<td>0 \ldots 1</td>
<td>( 1 - \sum_{x} \left( P(x) \right)^2 )</td>
</tr>
<tr>
<td>( s )</td>
<td>support</td>
<td>0 \ldots 1</td>
<td>( P(A \cup B) )</td>
</tr>
<tr>
<td>( c )</td>
<td>confidence</td>
<td>0 \ldots 1</td>
<td>( \min(P(A \cup B) - P(A), P(A \cup B) - P(B)) )</td>
</tr>
<tr>
<td>( L )</td>
<td>Laplace</td>
<td>0 \ldots 1</td>
<td>( \min(P(A \cup B) - P(A), P(A \cup B) - P(B)) )</td>
</tr>
<tr>
<td>( IS )</td>
<td>Cosine</td>
<td>0 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>coherence (acket)</td>
<td>0 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>all-confidence</td>
<td>0 \ldots 1</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{\min(P(A \cup B) - P(A), P(A \cup B) - P(B))} )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>odds ratio</td>
<td>0 \ldots \infty</td>
<td>( \frac{P(A \mid B)}{P(A^c \mid B)} \times \frac{P(B \mid A)}{P(B^c \mid A)} )</td>
</tr>
<tr>
<td>( V )</td>
<td>Correction</td>
<td>0.5 \ldots \infty</td>
<td>( \frac{P(A \cup B) - P(A) - P(B) + 1}{\min(P(A \cup B) - P(A), P(A \cup B) - P(B))} )</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>lift</td>
<td>0 \ldots \infty</td>
<td>( \frac{P(A \cup B) - P(A)P(B)}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
<tr>
<td>( S )</td>
<td>Collectivity</td>
<td>0 \ldots \infty</td>
<td>( \frac{P(A \cup B) - P(A)P(B)}{\sqrt{P(A)P(B)P(A^c)P(B^c)}} )</td>
</tr>
</tbody>
</table>

Simpson’s Paradox

E.g., kidney stone treatment

<table>
<thead>
<tr>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>78% (273 / 350)</td>
<td>83% (289 / 350)</td>
</tr>
</tbody>
</table>

**Simpson’s Paradox**

<table>
<thead>
<tr>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>78% (273 / 350)</td>
<td>83% (289 / 350)</td>
</tr>
</tbody>
</table>

Problem of *confounding variables*.

<table>
<thead>
<tr>
<th>small stones</th>
<th>large stones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment A</strong></td>
<td><strong>Treatment B</strong></td>
</tr>
<tr>
<td>93% (81 / 87)</td>
<td>87% (234 / 270)</td>
</tr>
</tbody>
</table>

---

**Chapter 5: Mining Frequent Patterns, Association and Correlations**

- Basic concepts and a road map
- Efficient and scalable frequent itemset mining methods
- Mining various kinds of association rules
- From association mining to correlation analysis
- Constraint-based association mining
- Summary
Frequent-Pattern Mining: Summary

- Frequent pattern mining—an important task in data mining
- Scalable frequent pattern mining methods
  - Apriori (Candidate generation & test)
  - Projection-based (FPgrowth, CLOSET+, ...)
  - Vertical format approach (CHARM, ...)
- Mining a variety of rules and interesting patterns
- Constraint-based mining
- Mining sequential and structured patterns
- Extensions and applications

Frequent-Pattern Mining: Research Problems

- Mining fault-tolerant frequent, sequential and structured patterns
  - Patterns allows limited faults (insertion, deletion, mutation)
- Mining truly interesting patterns
  - Surprising, novel, concise, ...
- Application exploration
  - E.g., DNA sequence analysis and bio-pattern classification
  - “Invisible” data mining
Ref: Basic Concepts of Frequent Pattern Mining

- **(Max-pattern)** R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98.
- **(Sequential pattern)** R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95

Ref: Apriori and Its Improvements

Ref: Depth-First, Projection-Based FP Mining

- J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. SIGMOD' 00.
- J. Pei, J. Han, and R. Mao. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets. DMKD'00.
- J. Liu, Y. Pan, K. Wang, and J. Han. Mining Frequent Item Sets by Opportunistic Projection. KDD'02.
- J. Han, J. Wang, Y. Lu, and P. Tzvetkov. Mining Top-K Frequent Closed Patterns without Minimum Support. ICDM'02.
- J. Wang, J. Han, and J. Pei. CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. KDD'03.

Ref: Vertical Format and Row Enumeration Methods

- Zaki and Hsiao. CHARM: An Efficient Algorithm for Closed Itemset Mining, SDM'02.
Ref: Mining Multi-Level and Quantitative Rules

- J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. VLDB'95.

Ref: Mining Correlations and Interesting Rules

- P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. KDD'02.
- E. Omiecinski. Alternative Interest Measures for Mining Associations. TKDE'03.
- Y. K. Lee, W.Y. Kim, Y. D. Cai, and J. Han. CoMine: Efficient Mining of Correlated Patterns. ICDM'03.
Ref: Mining Other Kinds of Rules

- K. Wang, S. Zhou, J. Han. Profit Mining: From Patterns to Actions. EDBT'02.

Ref: Constraint-Based Pattern Mining

- R. Ng, L.V.S. Lakshmanan, J. Han & A. Pang. Exploratory mining and pruning optimizations of constrained association rules. SIGMOD'98.
- J. Pei, J. Han, and L. V. S. Lakshmanan. Mining Frequent Itemsets with Convertible Constraints. ICDE'01.
- J. Pei, J. Han, and W. Wang, Mining Sequential Patterns with Constraints in Large Databases, CIKM'02.
Ref: Mining Sequential and Structured Patterns

- J. Pei, J. Han, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. ICDE'01.
- M. Kuramochi and G. Karypis. Frequent Subgraph Discovery. ICDM'01.
- X. Yan, J. Han, and R. Afshar. CloSpan: Mining Closed Sequential Patterns in Large Datasets. SDM'03.
- X. Yan and J. Han. CloseGraph: Mining Closed Frequent Graph Patterns. KDD'03.

Ref: Mining Spatial, Multimedia, and Web Data

- K. Koperski and J. Han, Discovery of Spatial Association Rules in Geographic Information Databases, SSD'95.
- O. R. Zaiane, M. Xin, J. Han, Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. ADL'98.
- O. R. Zaiane, J. Han, and H. Zhu, Mining Recurrent Items in Multimedia with Progressive Resolution Refinement. ICDE'00.
- D. Gunopulos and I. Tsoukas. Efficient Mining of Spatiotemporal Patterns. STD'01.
Ref: Mining Frequent Patterns in Time-Series Data

- B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. ICDE'98.
- J. Han, G. Dong and Y. Yin, Efficient Mining of Partial Periodic Patterns in Time Series Database, ICDE'99.
- H. Lu, L. Feng, and J. Han. Beyond Intra-Transaction Association Analysis: Mining Multi-Dimensional Inter-Transaction Association Rules. TOIS:00.
- B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online Data Mining for Co-Evolving Time Sequences. ICDE'00.

Ref: Iceberg Cube and Cube Computation

- Y. Zhao, P. M. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multi-dimensional aggregates. SIGMOD'97.
- S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of OLAP data cubes. EDBT'98.
Ref: Iceberg Cube and Cube Exploration

- J. Han, J. Pei, G. Dong, and K. Wang, Computing Iceberg Data Cubes with Complex Measures. SIGMOD’01.
- G. Dong, J. Han, J. Lam, J. Pei, and K. Wang. Mining Multi-Dimensional Constrained Gradients in Data Cubes. VLDB’01.
- L. V. S. Lakshmanan, J. Pei, and J. Han. Quotient Cube: How to Summarize the Semantics of a Data Cube. VLDB’02.
- D. Xin, J. Han, X. Li, B. W. Wah. Star-Cubing: Computing Iceberg Cubes by Top-Down and Bottom-Up Integration. VLDB’03.

Ref: FP for Classification and Clustering

- B. Liu, W. Hsu, Y. Ma. Integrating Classification and Association Rule Mining. KDD’98.
- W. Li, J. Han, and J. Pei. CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. ICDM’01.
- J. Yang and W. Wang. CLUSEQ: efficient and effective sequence clustering. ICDE’03.
- B. Fung, K. Wang, and M. Ester. Large Hierarchical Document Clustering Using Frequent Itemset. SDM’03.
- X. Yin and J. Han. CPAR: Classification based on Predictive Association Rules. SDM’03.
Ref: Stream and Privacy-Preserving FP Mining

- G. Manku and R. Motwani. Approximate Frequency Counts over Data Streams. VLDB'02.
- Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-Dimensional Regression Analysis of Time-Series Data Streams. VLDB'02.
- C. Giannella, J. Han, J. Pei, X. Yan and P. S. Yu. Mining Frequent Patterns in Data Streams at Multiple Time Granularities, Next Generation Data Mining:03.

Ref: Other Freq. Pattern Mining Applications

- T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining Database Structure; or How to Build a Data Quality Browser. SIGMOD'02.