Descriptive data analysis

- Aims to summarise the main qualitative traits of data.
- Used mainly for discovering underlying processes and relations in data.
- Facts are presented together with understandable explanations.
- Used for example in medical, physical and social sciences.

Example

```
23168
6538
542
1386
561
482
53624
```

E.g. set of ‘items’

```
Anne Britt Margus Gert
Gert Toomas Anne
Britt Gert Anne Laura
Toomas Gert Laura Britt
Laura Gert Anne Britt
Margus Britt Laura
```

Example

```
23168
6538
542
1386
561
482
53624
```

Example

```
12368
3568
245
1368
156
248
23456
```
Data Mining
Association Analysis: Basic Concepts and Algorithms
Lecture Notes for Chapter 6

Introduction to Data Mining
by
Tan, Steinbach, Kumar

Chapter 6:

Association Rule Mining
Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Market-Basket transactions
- Bread, Milk
- Bread, Diaper, Beer, Eggs
- Milk, Diaper, Beer, Coke
- Bread, Diaper, Beer
- Bread, Milk, Diaper, Coke

Example of Association Rules
(Diaper) → (Beer), (Milk, Bread) → (Eggs, Coke), (Beer, Bread) → (Milk).

Implication means co-occurrence, not causality!
Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
  - Example: {Milk, Bread, Diaper}
- **Support count (\(a\))**
  - Frequency of occurrence of an itemset
  - Example: \(\sigma\)({Milk, Bread, Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - Example: \(s\)({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a \(\text{minsup}\) threshold

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diaper, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diaper, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diaper, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diaper, Coke</td>
</tr>
</tbody>
</table>

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form \(X \rightarrow Y\), where \(X\) and \(Y\) are itemsets
  - Example: \(\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}\)
- **Rule Evaluation Metrics**
  - **Support (\(s\))**
    - Fraction of transactions that contain both \(X\) and \(Y\)
  - **Confidence (\(c\))**
    - Measures how often items in \(Y\) appear in transactions that contain \(X\)

\[
\begin{align*}
s &= \frac{\sigma(X \cup Y)}{|T|} \\
c &= \frac{\sigma(X \cup Y)}{\sigma(X)}
\end{align*}
\]

Example: \(\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}\)

\[
\begin{align*}
s &= \frac{2}{5} = 0.4 \\
c &= \frac{2}{0.67}
\end{align*}
\]

Association Rule Mining Task

- Given a set of transactions \(T\), the goal of association rule mining is to find all rules having
  - support \(\geq \text{minsup}\) threshold
  - confidence \(\geq \text{minconf}\) threshold
- **Brute-force approach**:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the \(\text{minsup}\) and \(\text{minconf}\) thresholds
  - \(\Rightarrow\) Computationally prohibitive!

Mining Association Rules

- **Example of Rules**:
  - \(\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}\) (\(s=0.4, c=0.67\))
  - \(\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}\) (\(s=0.4, c=0.5\))

- **Observations**:
  - All the above rules are binary partitions of the same itemset: \(\{\text{Milk, Diaper, Beer}\}\)
  - Rules originating from the same itemset have identical support but can have different confidence
  - Thus, we may decouple the support and confidence requirements

Mining Association Rules

- **Two-step approach**:
  1. **Frequent Itemset Generation**
     - Generate all itemsets whose support \(\geq \text{minsup}\)
  2. **Rule Generation**
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
- **Frequent itemset generation is still computationally expensive**
Frequent Itemset Generation

Given d items, there are $2^d$ possible candidate itemsets.

Brute-force approach:
- Each itemset in the lattice is a candidate frequent itemset
- Count the support of each candidate by scanning the database
- Match each transaction against every candidate
- Complexity ~ $O(NMw)$ => Expensive since $M = 2^d$ !!!

Example – counting frequent substrings
- How many “words” of length k, alphabet $\Sigma$?
- How to count the frequency of every possible substring in protein sequences?
- How to count all words in a book?
- Substring vs regular expressions

Computational Complexity
- Given d unique items:
  - Total number of itemsets = $2^d$
  - Total number of possible association rules:
    $$ R = \sum_{k=1}^{d} \left[ \binom{d}{k} \times \sum_{j=0}^{k} \binom{d-k}{j} \right] $$
    $$ = 3^d - 2^d + 1 $$
    If d=6, R = 602 rules

Frequent Itemset Generation Strategies
- Reduce the number of candidates (M)
  - Complete search: $M = 2^d$
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by DHP and vertical-based mining algorithms
- Reduce the number of comparisons (NM)
  - Use efficient data structures to store the candidates or transactions
  - No need to match every candidate against every transaction

Reducing Number of Candidates
- Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:
  $$ \forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y) $$
  - Support of an itemset never exceeds the support of its subsets
  - This is known as the anti-monotone property of support
Apriori Algorithm

- Method:
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

Generate Hash Tree

Suppose you have 15 candidate itemsets of length 3:

- {1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {2 3 6}, {3 6 8}

You need:
- Hash function
  - Max leaf size: max number of itemsets stored in a leaf node (if number of candidate itemsets exceeds max leaf size, split the node)

Reducing Number of Comparisons

- Candidate counting:
  - Scan the database of transactions to determine the support of each candidate itemset
  - To reduce the number of comparisons, store the candidates in a hash structure
    - Instead of matching each transaction against every candidate, match it against candidates contained in the hashed buckets

Association Rule Discovery: Hash tree

Hash function

- Hash on 1, 4 or 7
Given a transaction $t$, what are the possible subsets of size 3?

Subset Operation

Subset Operation Using Hash Tree
**Factors Affecting Complexity**

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - more space is needed to store support count of each item
  - if number of frequent items also increases, both computation and I/O costs may also increase
- Size of database
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions
- Average transaction width
  - transaction width increases with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)

**Compact Representation of Frequent Itemsets**

- Some itemsets are redundant because they have identical support as their supersets
- Number of frequent itemsets = \(3 \times \frac{10!}{5!} \)
- Need a compact representation

**Maximal Frequent Itemset**

An itemset is maximal frequent if none of its immediate supersets is frequent

**Closed Itemset**

- An itemset is closed if none of its immediate supersets has the same support as the itemset
Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - General-to-specific vs Specific-to-general

Maximal Frequent Itemset

An itemset is maximal frequent if none of its immediate supersets is frequent.

Maximal vs Closed Itemsets

Maximal vs Closed Frequent Itemsets

- Maximum Frequent Itemset

- Frequent Itemsets
- Closed Frequent Itemsets
- Maximal Frequent Itemsets

Maximal Frequent Itemset

- Maximum Frequent Itemset

Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - General-to-specific vs Specific-to-general

Maximal Frequent Itemset
Alternative Methods for Frequent Itemset Generation

- Traversal of Itemset Lattice
  - Breadth-first vs Depth-first

FP-growth Algorithm

- Use a compressed representation of the database using an FP-tree
- Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets
- Frequent pattern growth
  - Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery 8:53-87, 2004.

Why FP, not Apriori?

Apriori works well except when:
- Lots of frequent patterns
  - Big set of items
  - Low minimum support threshold
  - Long patterns

Why: Candidate sets become huge
- \(10^4\) frequent patterns of length 1 \(\rightarrow\) \(10^8\) length 2 candidates
- Discovering pattern of length 100 requires at least \(2^{100}\) candidates (nr of subsets)
- Repeated database scans costly (long patterns)

FP-tree construction
Figure 6.25. An FP-tree representation for the data set shown in Figure 6.24 with a different item ordering scheme.

FP-Tree Construction

Figure 6.26. Decomposing the frequent itemset generation problem into multiple subproblems, where each subproblem involves finding frequent itemsets ending in c, d, e, and a.

Figure 6.27. Example of applying the FP-growth algorithm to find frequent itemsets ending in e.

FP-growth

Conditional Pattern base for D:
P = {(A:1,B:1,C:1), (A:1,B:1), (A:1,C:1), (A:1), (B:1,C:1)}
Recursively apply FP-growth on P
Frequent Itemsets found (with sup > 1):
AD, BD, CD, ACD, BCD

All frequent itemsets of this example

E: {e}, {d,e}, {a,d,e}, {c,e}, {a,e}
D: {d}, {c,d}, {b,c,d}, {a,c,d}, {b,d}, {a,b,d}, {a,d}
C: {c}, {b,c}, {a,b,c}, {a,c}
B: {b}, {a,b}
A: {a}
Tree Projection

- Items are listed in lexicographic order
- Each node $P$ stores the following information:
  - Itemset for node $P$
  - List of possible lexicographic extensions of $P$: $E(P)$
  - Pointer to projected database of its ancestor node
  - Bitvector containing information about which transactions in the projected database contain the itemset

Projected Database

Original Database:  
<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(A,B)</td>
</tr>
<tr>
<td>2</td>
<td>(B,C,D)</td>
</tr>
<tr>
<td>3</td>
<td>(A,C,D,E)</td>
</tr>
<tr>
<td>4</td>
<td>(A,D,E)</td>
</tr>
<tr>
<td>5</td>
<td>(A,B,C)</td>
</tr>
<tr>
<td>6</td>
<td>(A,B,C,D)</td>
</tr>
<tr>
<td>7</td>
<td>(B,C)</td>
</tr>
<tr>
<td>8</td>
<td>(A,B,C)</td>
</tr>
<tr>
<td>9</td>
<td>(A,B,D)</td>
</tr>
<tr>
<td>10</td>
<td>(B,C,D,E)</td>
</tr>
</tbody>
</table>

Projected Database for node $A$:

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(B)</td>
</tr>
<tr>
<td>2</td>
<td>(B,C)</td>
</tr>
<tr>
<td>3</td>
<td>(C,D,E)</td>
</tr>
<tr>
<td>4</td>
<td>(D,E)</td>
</tr>
<tr>
<td>5</td>
<td>(B,C)</td>
</tr>
<tr>
<td>6</td>
<td>(B,C)</td>
</tr>
<tr>
<td>7</td>
<td>(B,C)</td>
</tr>
<tr>
<td>8</td>
<td>(B,C)</td>
</tr>
<tr>
<td>9</td>
<td>(B,D)</td>
</tr>
<tr>
<td>10</td>
<td>(B,D)</td>
</tr>
</tbody>
</table>

For each transaction $T$, projected transaction at node $A$ is $T \cap E(A)$

ECLAT

- For each item, store a list of transaction ids (tids)

Horizontal Data Layout

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A,B,E</td>
</tr>
<tr>
<td>2</td>
<td>B,C,D</td>
</tr>
<tr>
<td>3</td>
<td>C,E</td>
</tr>
<tr>
<td>4</td>
<td>A,C,D</td>
</tr>
<tr>
<td>5</td>
<td>A,B,C,D</td>
</tr>
<tr>
<td>6</td>
<td>A,E</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>A,B,C</td>
</tr>
<tr>
<td>9</td>
<td>A,C,D</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
</tr>
</tbody>
</table>

Vertical Data Layout

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TID-list

ECLAT

- Determine support of any $k$-itemset by intersecting tid-lists of two of its $(k-1)$ subsets.
- 3 traversal approaches:
  - top-down, bottom-up and hybrid
  - Advantage: very fast support counting
  - Disadvantage: intermediate tid-lists may become too large for memory
Rule Generation

- Given a frequent itemset \( L \), find all non-empty subsets \( f \subseteq L \) such that \( f \rightarrow L - f \) satisfies the minimum confidence requirement.
  - If \( \{A, B, C, D\} \) is a frequent itemset, candidate rules:
    - \( ABC \rightarrow D \), \( ABD \rightarrow C \), \( ACD \rightarrow B \), \( BCD \rightarrow A \), \( AB \rightarrow CD \), \( AC \rightarrow BD \), \( AD \rightarrow BC \), \( BC \rightarrow AD \), \( BD \rightarrow AC \), \( CD \rightarrow AB \).
- If \( |L| = k \), then there are \( 2^k - 2 \) candidate association rules (ignoring \( L \rightarrow \emptyset \) and \( \emptyset \rightarrow L \)).

Rule Generation for Apriori Algorithm

- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent.
  - \( \text{join}(CD \Rightarrow AB, BD \Rightarrow AC) \) would produce the candidate rule \( D \Rightarrow ABC \).
- Prune rule \( D \Rightarrow ABC \) if its subset \( AD \Rightarrow BC \) does not have high confidence.

Effect of Support Distribution

- Many real data sets have skewed support distribution.

Effect of Support Distribution

- How to set the appropriate \( \text{minsup} \) threshold?
  - If \( \text{minsup} \) is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products).
  - If \( \text{minsup} \) is set too low, it is computationally expensive and the number of itemsets is very large.
- Using a single minimum support threshold may not be effective.
Multiple Minimum Support

How to apply multiple minimum supports?
- MS(i): minimum support for item i
  - e.g.: MS(Milk) = 5%, MS(Broccoli) = 0.1%, MS(Salmon) = 0.5%
  - MS(Milk, Broccoli) = min (MS(Milk), MS(Broccoli)) = 0.1%

- Challenge: Support is no longer anti-monotone
  - Suppose: Support(Milk, Coke) = 1.5% and Support(Milk, Coke, Broccoli) = 0.5%
    - {Milk, Coke} is infrequent but {Milk, Coke, Broccoli} is frequent

---

Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if (A, B, C) → (D) and (A, B) → (D) have same support & confidence

- Interestingness measures can be used to prune/rank the derived patterns

- In the original formulation of association rules, support & confidence are the only measures used
Application of Interestingness Measure

Interestingness Measures

Knowledge

Preprocessing Data

Postprocessing

Mining

Preprocessing

Selected Data

Data

Laylines and patterns in data ...

Stores built according to "mystical" plan

137 random points, 80 4-point alignments

Evaluation of Associated Patterns

- Objective measure of interestingness (statistics)
- Visualization
- Template based approaches
  - Report those matching user templates
- Subjective interestingness
  - Domain information, concept hierarchy, profit margins, etc
Computing Interestingness Measure

Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table.

Contingency table for $X \rightarrow Y$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$\neg Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$f_{11}$</td>
<td>$f_{10}$</td>
</tr>
<tr>
<td>$\neg X$</td>
<td>$f_{01}$</td>
<td>$f_{00}$</td>
</tr>
<tr>
<td>$f_{+1}$</td>
<td>$f_{+0}$</td>
<td></td>
</tr>
</tbody>
</table>

$T$  

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

Diaper, Milk $\rightarrow$ Beer

- support $400/3400$, confidence $300/400$

<table>
<thead>
<tr>
<th>Beer</th>
<th>Not Beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diaper, Milk</td>
<td>300</td>
</tr>
<tr>
<td>Not Diaper, Milk</td>
<td>1000</td>
</tr>
</tbody>
</table>

Diaper, Milk $\rightarrow$ Beer

- support $400/3400$, confidence $300/400$

<table>
<thead>
<tr>
<th>Beer</th>
<th>Not Beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diaper, Milk</td>
<td>300</td>
</tr>
<tr>
<td>Not Diaper, Milk</td>
<td>1000</td>
</tr>
</tbody>
</table>

Drawback of Confidence

<table>
<thead>
<tr>
<th>Coffee</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Association Rule: Tea $\rightarrow$ Coffee

Confidence = $P($Coffee$|$Tea) = 0.75

but $P($Coffee$)$ = 0.9

- Although confidence is high, rule is misleading
- $P($Coffee$|$Tea$)$ = 0.9375

Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)

- $P($S and B$) = 420/1000 = 0.42$
- $P($S$) \times P($B$) = 0.6 \times 0.7 = 0.42$
- $P($S and B$) = P($S$) \times P($B$) \Rightarrow$ Statistical independence
- $P($S and B$) > P($S$) \times P($B$) \Rightarrow$ Positively correlated
- $P($S and B$) < P($S$) \times P($B$) \Rightarrow$ Negatively correlated
Statistical-based Measures

- Measures that take into account statistical dependence

\[
\text{Lift} = \frac{P(Y|X)}{P(Y)}
\]

\[
\text{Interest} = \frac{P(X,Y)}{P(X)P(Y)}
\]

\[
\text{PS} = P(X,Y) - P(X)P(Y)
\]

\[
\phi - \text{coefficient} = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)\left[1 - P(X)P(Y)\right]}}
\]

Example: Lift/Interest

<table>
<thead>
<tr>
<th></th>
<th>Coffee</th>
<th>Coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tea</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Tea</td>
<td>75</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Association Rule: Tea → Coffee

Confidence = \( P(\text{Coffee}|\text{Tea}) = 0.75 \)

but \( P(\text{Coffee}) = 0.9 \)

→ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)

Diaper, Milk -> Beer

- support 400/3400, confidence 300/400

→ Lift (300/400) / (1300/3400) = 1.96

<table>
<thead>
<tr>
<th></th>
<th>Beer</th>
<th>Not Beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diaper,Milk</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>Not Diaper,Milk</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>1300</td>
<td>2100</td>
</tr>
</tbody>
</table>

Example

\[
P(X \mid Y) = \frac{P(X,Y)}{P(Y)}
\]

\[
P(Y \mid X) = \frac{P(X,Y)}{P(X)}
\]

\[
P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}
\]

Drawback of Lift & Interest

\[
\text{Lift} = \frac{0.1}{(0.1)(0.1)} = 10
\]

\[
\text{Lift} = \frac{0.9}{(0.9)(0.9)} = 1.11
\]

Statistical independence:

If \( P(X,Y) = P(X)P(Y) \) => Interest = 0.9

\[
\text{Lift} = \frac{0.1}{(0.1)(0.1)} = 10
\]

\[
\text{Lift} = \frac{0.9}{(0.9)(0.9)} = 1.11
\]
Confounding factors

- There may be some other hidden factors
- Stratified data may show different results
- Statistics …

### Table 6.12. Examples of asymmetric objective measures for the rule $A \rightarrow B$.

<table>
<thead>
<tr>
<th>Measure (Symbol)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodman-Kruskal ($\lambda$)</td>
<td>$\lambda = \frac{\sum_{i} \max(f_{AB} - f_{A}f_{B})/(N - f_{A}f_{B})}{N}$</td>
</tr>
<tr>
<td>Mutual Information ($M$)</td>
<td>$M = \sum_{i} \frac{f_{ij}}{N} \log \frac{N f_{ij}}{f_{i} f_{j}}$</td>
</tr>
<tr>
<td>J-Measure ($J$)</td>
<td>$J = \frac{f_{ij}/N}{(f_{i} + 1)/(f_{i} + 2)}$</td>
</tr>
<tr>
<td>Gini index ($G$)</td>
<td>$G = \frac{1}{N} \sum_{i,j} \left( \frac{f_{ij}}{N} - \frac{f_{i}}{N} \frac{f_{j}}{N} \right)^2$</td>
</tr>
<tr>
<td>Conviction ($V$)</td>
<td>$V = \frac{f_{ij}/N}{(f_{i} + 1)/(f_{i} + 2)}$</td>
</tr>
<tr>
<td>Certainty factor ($F$)</td>
<td>$F = \frac{f_{ij}/N}{(1 - f_{ij}/N)}$</td>
</tr>
<tr>
<td>Added Value ($AV$)</td>
<td>$AV = f_{ij}/N$</td>
</tr>
</tbody>
</table>

### Table 6.11. Examples of symmetric objective measures for the itemset $\{A, B\}$.

<table>
<thead>
<tr>
<th>Measure (Symbol)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation ($\rho$)</td>
<td>$\rho = \frac{N f_{AB} f_{BA} - f_{A} f_{B} f_{AB} f_{BA}}{\sqrt{(N f_{A} f_{AB} f_{BA} f_{AB} - f_{A} f_{B} f_{AB} f_{BA})(N f_{A} f_{B} f_{AB} f_{BA} - f_{A} f_{B} f_{AB} f_{BA})}}$</td>
</tr>
<tr>
<td>Odds ratio ($\alpha$)</td>
<td>$\alpha = \frac{f_{AB} f_{BA}}{f_{A} f_{B}}$</td>
</tr>
<tr>
<td>Kappa ($\kappa$)</td>
<td>$\kappa = \frac{N f_{AB} f_{BA} - f_{A} f_{B} f_{AB} f_{BA} + f_{A} f_{B} f_{AB} f_{BA}}{N f_{A} f_{B} f_{AB} f_{BA} - f_{A} f_{B} f_{AB} f_{BA}}$</td>
</tr>
<tr>
<td>Interest ($I$)</td>
<td>$I = \frac{(N f_{AB} f_{BA} - f_{A} f_{B} f_{AB} f_{BA})^2}{f_{A} f_{B} f_{AB} f_{BA}}$</td>
</tr>
<tr>
<td>Cosine ($\cos$)</td>
<td>$\cos = \frac{f_{AB} f_{BA}}{f_{A} f_{B}}$</td>
</tr>
<tr>
<td>Pratt-Shapiro ($PS$)</td>
<td>$PS = \frac{N f_{AB} f_{BA}}{f_{A} f_{B}}$</td>
</tr>
<tr>
<td>Collective strength ($S$)</td>
<td>$S = \frac{N f_{AB} f_{BA}}{f_{A} f_{B}}$</td>
</tr>
<tr>
<td>Jaccard ($\zeta$)</td>
<td>$\zeta = \frac{f_{AB}}{f_{A} + f_{B} - f_{AB}}$</td>
</tr>
<tr>
<td>All-confidence ($\lambda$)</td>
<td>$\lambda = \min \left( \frac{f_{AB}}{f_{A}}, \frac{f_{BA}}{f_{B}} \right)$</td>
</tr>
</tbody>
</table>

There are lots of measures proposed in the literature, some measures are good for certain applications, but not for others. What criteria should we use to determine whether a measure is good or bad? What about Apriori-style support-based pruning? How does it affect these measures?
Properties of A Good Measure

Piatetsky-Shapiro:

3 properties a good measure M must satisfy:

- \( M(A, B) = 0 \) if A and B are statistically independent
- \( M(A, B) \) increase monotonically with \( P(A, B) \) when \( P(A) \) and \( P(B) \) remain unchanged
- \( M(A, B) \) decreases monotonically with \( P(A) \) [or \( P(B) \)] when \( P(A, B) \) and \( P(B) \) [or \( P(A) \)] remain unchanged

Comparing Different Measures

Example:

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Rankings of contingency tables using various measures:

![Table comparing different measures]

Property under Variable Permutation

Does \( M(A, B) = M(B, A) \)?

Symmetric measures:
- support, lift, collective strength, cosine, Jaccard, etc

Asymmetric measures:
- confidence, conviction, Laplace, J-measure, etc

Grade-Gender Example (Mosteller, 1968):

Mosteller:
Underlying association should be independent of the relative number of male and female students in the samples

![Table showing grade-gender example]

Property under Inversion Operation

Example: \( \phi \)-Coefficient

- \( \phi \)-coefficient is analogous to correlation coefficient for continuous variables

\[
\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238
\]

\[
\phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238
\]

Coefficient is the same for both tables
Property under Null Addition

Invariant measures:
- support, cosine, Jaccard, etc

Non-invariant measures:
- correlation, Gini, mutual information, odds ratio, etc

Support-based Pruning

- Most of the association rule mining algorithms use support measure to prune rules and itemsets.
- Study effect of support pruning on correlation of itemsets:
  - Generate 10000 random contingency tables
  - Compute support and pairwise correlation for each table
  - Apply support-based pruning and examine the tables that are removed.

Effect of Support-based Pruning

Support-based pruning eliminates mostly negatively correlated itemsets.

Different Measures have Different Properties

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Measure</th>
<th>Range</th>
<th>PI</th>
<th>PF</th>
<th>OF</th>
<th>GS</th>
<th>SD</th>
<th>GD</th>
<th>DS</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Correlation</td>
<td>-1...1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>R</td>
<td>Odds ratio</td>
<td>0...1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Y</td>
<td>Yule's Y</td>
<td>-1...0...1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>K</td>
<td>Gini</td>
<td>0...1</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>J</td>
<td>J-Measure</td>
<td>0...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>S</td>
<td>Support</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>C</td>
<td>Centrality</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>L</td>
<td>Laplace</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>V'</td>
<td>Correlation</td>
<td>0.5...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>M</td>
<td>Mutual_info</td>
<td>0...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>G</td>
<td>Gini</td>
<td>0...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>a</td>
<td>Support</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>b</td>
<td>Centrality</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>c</td>
<td>Laplace</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>V</td>
<td>Correlation</td>
<td>0.5...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>J</td>
<td>J-Measure</td>
<td>0...1</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>S</td>
<td>Support</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>C</td>
<td>Centrality</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>L</td>
<td>Laplace</td>
<td>0...1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Effect of Support-based Pruning

- Investigate how support-based pruning affects other measures.
- Steps:
  - Generate 10000 contingency tables
  - Rank each table according to the different measures
  - Compute the pair-wise correlation between the measures.
Subjective Interestingness Measure

- **Objective measure:**
  - Rank patterns based on statistics computed from data
  - e.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc.).

- **Subjective measure:**
  - Rank patterns according to user’s interpretation
    - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is actionable (Silberschatz & Tuzhilin)

Interestingness via Unexpectedness

- **Need to model expectation of users (domain knowledge)**
- **Need to combine expectation of users with evidence from data (i.e., extracted patterns)**

Web Data (Cooley et al 2001)

- Domain knowledge in the form of site structure
  - Given an itemset $F = \{X_1, X_2, \ldots, X_k\}$ ($X_i$: Web pages)
  - $L$: number of links connecting the pages
  - $lfactor = L / (k \times k - 1)$
  - $cfactor = 1$ (if graph is connected), 0 (disconnected graph)
- Structure evidence = $cfactor \times lfactor$
- Usage evidence $= \frac{P(X_1 \cup X_2 \cup \ldots \cup X_k)}{P(X_1, X_2, \ldots, X_k)}$
- Use Dempster-Shafer theory to combine domain knowledge and evidence from data
Figure 4.3.1: A summary of the various research activities in association analysis.

## Additional links

- [http://www.kdnuggets.com/](http://www.kdnuggets.com/)