Fast algorithm for mining association rules

Presenter: Linda
Discusser: Massih
Instructor: Rachel Pottinger
University of British Columbia

Overview

• Introduction of data mining (application of mining)
• Association rule mining
• Apriori algorithm
• Comparison of different algorithms
• Performance comparison

Introduction of data mining

• What is data mining?
Data mining (DM) is the process of automatically searching large volumes of data for patterns such as association rules.

• A Real-World Example:
a supermarket chain, through analysis of transactions over a long period of time, found that beer and diapers were often bought together

Introduction of data mining

What are the uses of data mining?

• Market segmentation - Identify common characteristics of customers who buy the same products from your company.
• Customer churn - Predict which customers are likely to leave and go to a competitor.
• Fraud detection - Identify which transactions are likely to be fraudulent.
• Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
• Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
• Market basket analysis - Understand what products or services are commonly purchased together; e.g., beer and diapers.
• Trend analysis - Reveal the difference between a typical customer this month and last.

Introduction of data mining

• How does data mining work?
Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

• What types of relationships are sought?
  - **Classes**: Stored data is used to locate data in predetermined groups.
    example: a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
  - **Clusters**: Data items are grouped according to logical relationships or consumer preferences.
    example: data can be mined to identify market segments or consumer affinities.

Relationship Types (cont.)

• **Sequential patterns**: Data is mined to anticipate behavior patterns and trends.
  example: an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

• **Associations**: Data can be mined to identify associations.
The beer-diaper example is an example of associative mining.
**Association rule mining**

- **Itemset**
  - A set of itemsets is referred to as itemset
  - An itemset containing k items is called k-itemset

- **Associate rule**
  - \( X \Rightarrow Y \) (X and Y are disjoint sets of item set)
  - Support of an association rule is the percentage of the relevant data transaction for which the rule is true.

\[
\text{support}(A \Rightarrow B) = \frac{\text{# tuples containing both A and B}}{\text{total # of tuples}}
\]

- Confidence of an association is the measure of certainty associated with each discovered pattern.

\[
\text{confidence}(A \Rightarrow B) = \frac{\text{# tuples containing both A and B}}{\text{total # of tuples}}
\]

**Example**

- **Transaction database TDB**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c</td>
<td>3</td>
</tr>
<tr>
<td>d, e, f</td>
<td>2</td>
</tr>
<tr>
<td>g, h, i</td>
<td>1</td>
</tr>
<tr>
<td>j, k, l</td>
<td>2</td>
</tr>
<tr>
<td>m, n</td>
<td>1</td>
</tr>
</tbody>
</table>

**Problem decomposition**

- Discover large itemsets
- Use the largest itemsets to generate the desired rules

**Algorithm Apriori**

- Find large items

\( F_1 \) = \{ frequent 1-item sets \};

\( k = 2 \);

while( \( F_k \) is not empty ) {

\( C_k \) = \text{Apriori generate}( F_k );

for all transactions \( t \) in \( T \) {

\( \text{Subsets}(C_k, t) \);

\( F_{k+1} \) = \{ \text{c in } C_k \text{ s.t. c.count } \geq \text{ minimum support} \};

}

Answer = union of all sets \( F_k \);

**The Apriori algorithm-- Example**

- Find candidate pairs, count them
- Large itemsets of items
AprioriTid

- Uses the database only once.
- Builds a storage set $C^k$
  - Members has the form < TID, $\{X_k\}$ >
  - $X_k$ are potentially large $k$-items in transaction TID.
  - For $k=1$, $C^1$ is the database.
- Uses $C^k$ in pass $k+1$.

Example Of AprioriTid

Other algorithms

Apriori VS AprioriTid

- In the earlier passes, Apriori does better than AprioriTid.
- AprioriTid beats Apriori in later passes.

AprioriHybrid

- Uses Apriori in the initial passes and switches to AprioriTid when it expects that the set $C_k$ at the end of the pass will fit in memory.

Other algorithms

- AIS
- SETM
- Apriori and AprioriTid algorithms are much better than the AIS algorithm.
- AIS always did considerably better than SETM.
summary

- Association rules are an important tool in analyzing databases.
- We’ve seen an algorithm which finds all association rules in a database.
- The algorithm has better time results than previous algorithms.
- The algorithm maintains its performances for large databases.

Discussion

- What are challenges of data mining
- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distribution
- Streaming data