Multidimensional Pattern mining

Research on Pattern Mining

Pattern Mining-Road Map

Based on patten diversity pattern mining can be classified as: Basic patterns, based on abstraction levels, no of dimensions, types of values handled and based on criteria used to mine.

- **Basic Patterns**: Frequent pattern, closed , maximal pattern, infrequent and rare patterns.
- **Frequent pattern**: Pattern satisfies minimum support.
- **Closed pattern**: No super pattern p' with same support as p
- **Max-pattern:** If there exists no frequent superpattern of p
- **Rare patterns**: Patterns that occur rarely
- **Negative patterns**: Patterns have negative correlation.

Pattern Mining-Road Map

Closed pattern: A pattern X is closed in a dataset D if X is frequent and there exists no proper superitemset Y such that Y has same support as X in D.

{A}=4, {B}=2,{C}=5,{D}=4,{E}=6

{AB}=1,{AC}=3,{AD}=3,{AE}=4

A is closed due to {AE}

Maximal Itemset: A pattern X is not a max-pattern if X is frequent and if there exists at least one superset that is frequent.
D={4} {A,D}=3,{C,D}=3 {D,E}=3
D is closed but not maximal due to {A,D}

{ACE}=3 {ABCE}=1 maxiaml frequent

Rare Patterns vs. Negative Patterns

- Rare patterns
  - Very low support but interesting (e.g., buying Rolex watches)
  - How to mine them? Setting individualized, group-based min-support thresholds for different groups of items
- Negative patterns
  - Negatively correlated: Unlikely to happen together
  - Ex.: Since it is unlikely that the same customer buys both a Ford Expedition (an SUV car) and a Ford Fusion (a hybrid car), buying a Ford Expedition and buying a Ford Fusion are likely negatively correlated patterns

Pattern Mining-Road Map

Based on abstraction levels involved in a pattern:

- Patterns or association rules may have items or concepts residing at high, low, or multiple abstraction levels.
- Suppose a set of association rules mined includes the following rules where X is a variable representing a customer:

buys(X, “computer”)⇒buys(X, “printer”)

buys(X, “laptop computer”)⇒buys(X, “color laser printer”)

- The items in the rules when referenced at different abstraction levels called as **multilevel association rules**
- If the rules do not reference items or attributes at different abstraction levels, then called as **single-level association rules**

Pattern Mining-Road Map

Based on the number of dimensions involved in the rule or pattern:

- The rules can be **single-dimensional** association rules when the item or attribute refer to only one dimension Eg: buys
- If a rule/pattern references two or more dimensions, such as age, income, and buys it is a **multidimensional association rule/pattern**
- Example of a multidimensional rule:
- age(X, “20...29”)∧income(X, “52K … 58K”)⇒buys(X, “iPad”).

Pattern Mining-Road Map

Based on the types of values handled in the rule or pattern:

- If a rule involves associations between the presence or absence of items, it is a **Boolean association rule.** Eg: buys(X, computer")⇒buys(X, "printer")
- If a rule describes associations between quantitative items or attributes, then it is a **quantitative association rule**
- Eg: The quantitative attributes are age and income have been discretized within range

Based on the constraints or criteria used to mine selective patterns

- The patterns or rules when satisfying a set of user-defined constraints
- **Approximate,compressed,near-match**- the support count of the near or almost matching itemsets
- **Top-k** : the k most frequent itemsets for a user-specified value,
- **K redundancy-aware top-k**: The top-k patterns with similar or redundant patterns excluded and so on.

Pattern Mining-Road Map

Pattern mining can be classified with respect to kinds of data and applications involved

Based on kinds of data and features to be mined:

- Given relational and data warehouse data, most people are interested in frequent itemsets mining.
- **Sequence pattern:** The order in which items are frequently purchased(pc, camera and memory card).
- **Structural Patterns:** Frequent substructures in a structured data set

Based on application domain-specific semantics:

Both data and applications can be diverse patterns

Mining based on the domain specific semantics

Diversity lead to different pattern mining methodologies

Pattern Mining-Road Map

- **Based on data analysis usage:**
  - Pattern-based classification: Feature extraction step for classification
  - Pattern-based clustering: Clustering high dimensional data
  - Recommender system: Pattern used for semantic annotation or contextual analysis.

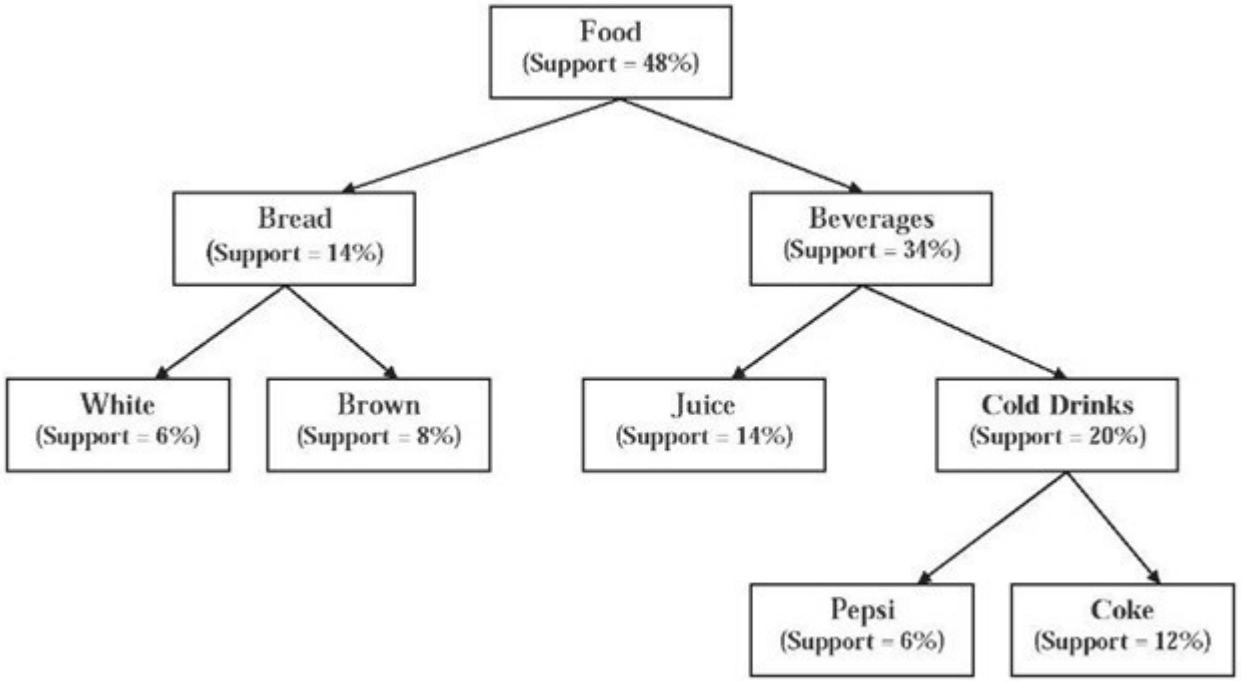
Mining Diverse Patterns

- Mining Multiple-Level Associations
- Mining Multi-Dimensional Associations
- Mining Quantitative Associations

Mining Multi level Associations

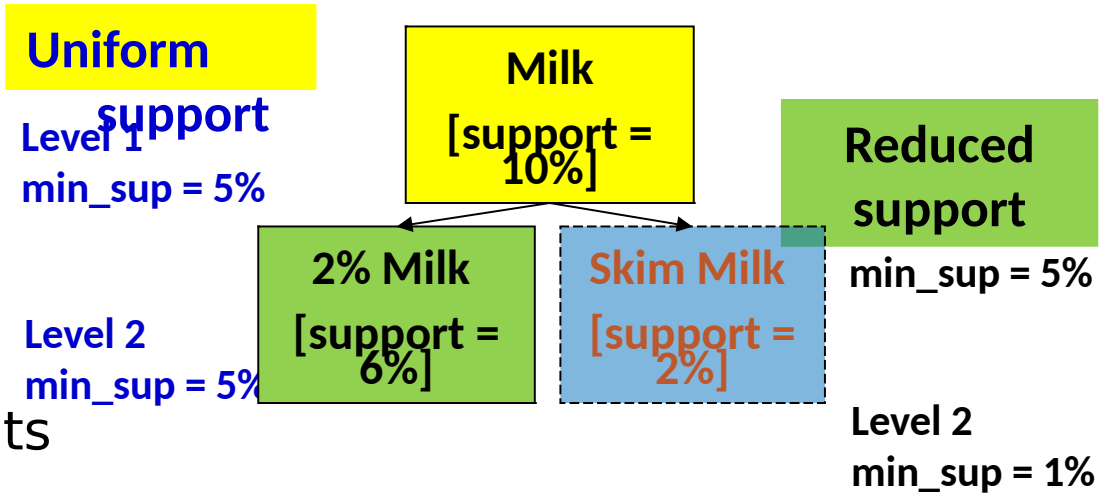
- Associations rules generated from mining data at multiple abstraction levels are called multiple-level or multilevel association rules.
- Multi level can be mined based on concept hierarchies under a support-confidence framework.
- Concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level more general concepts.
- A top-down strategy is employed at each concept level, at each level an algorithm for discovering frequent itemsets such as Apriori and its variations is used

Mining Multi level Associations

- 

Mining Multiple-Level Frequent Patterns (Uniform support)

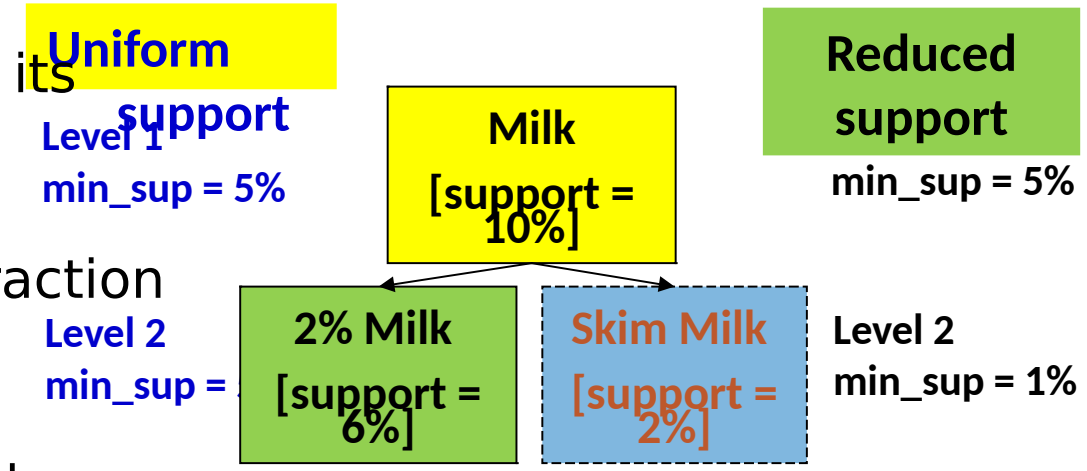
- Same minimum support is used at each abstraction levels



- Avoids examining itemsets whose ancestors do not have minimum support
- Threshold is high: Miss some meaningful associations occurring at low abstraction levels.
- Low threshold: Generates some uninteresting associations at high abstraction levels.

Mining Multiple-Level Frequent Patterns (Reduced Support)

- Each abstraction has its own minimum support
- The deeper the abstraction level the smaller corresponding threshold
- Almost all are considered as frequent



Multi-level Association: Flexible Support and Redundancy filtering

- Desirable to user-specific,item or group based minimal thresholds for mining multilevel rules.

Flexible min-support thresholds: Some items are more valuable but less frequent use group-based "individualized" min-support

  - Use non-uniform, group-based min-support
  - E.g., {diamond, watch, camera}: 0.05%; {bread, milk}: 5%; …

- Redundancy Filtering: Some rules may be redundant due to "ancestor" relationships between items
  - *Laptop computer => HP Printer [support = 8%, confidence = 70%]*
  - *Dell Laptop computer => HP Printer[support = 2%, confidence = 72%]*

  The first rule is an ancestor of the second rule
- A rule is redundant if its support is close to the "expected" value, based on the rule's ancestor
- Second rule is general rule doesn't not provide any new information pruned

Mining Multi-Dimensional Associations

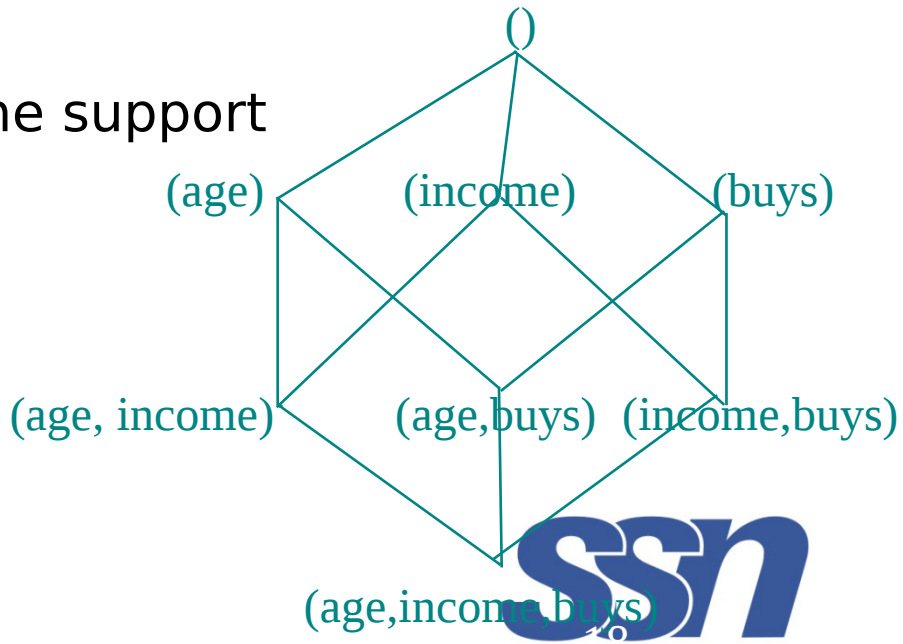
- Single-dimensional rules (Rule contain single distinct predicate)
  - buys(X, “milk”) ⇒ buys(X, “bread”)
- Multi-dimensional rules (i.e., items in ≥ 2 dimensions or predicates)
  - *no repeated predicates*
    - age(X, “18-25”) ∧ occupation(X, “student”) ⇒ buys(X, “coke”)
  - Hybrid-dimension association rules (*repeated predicates*)
    - age(X, “18-25”) ∧ buys(X, “popcorn”) ⇒ buys(X, “coke”)
- Attributes can be categorical or numerical
  - Categorical Attributes (e.g., *profession, product*: no ordering among values and infinite number of values): Data cube for inter-dimension association
  - Quantitative Attributes: Numeric, implicit ordering among values—discretization, clustering, and gradient approaches

Data cube – Based Mining of Quantitative Attributes

- Quantitative attributes discretized prior to mining using concept hierarchy.
- Numeric values are replaced by ranges
- Data cube is well suited for mining of multidimensional association rules with transformed multidimensional data
- Store aggregates in multidimensional space
- The cells of an n-dimensional

  cuboid correspond to store the support

  Counts of n-predicate sets
- Mining from data cubes can be much faster



Data cube – Based Mining of Quantitative Attributes

- Fetch the aggregate values and return the rules using rule aggregation algorithm
- Apriori property can be used to prune the search space
- If k-predicate has supp less than minimum support then the exploration of the predicate is terminated

Mining clustering based Quantitative Associations

- Quantitative association rules are generated using clustering
- Top-down Approach: Find the clusters using k-means or density based clustering algorithm
  - Examine the 2D spaces generated by combining the cluster with other cluster or nominal value of another dimension
  - Apriori property is applied if the support of the combination does not have minimum support
- Bottom down Approach: First clustering in high-dimensional space and then projecting in fewer dimensional spaces.