UNIT II

DATA PREPROCESSING ATTRIBUTES, & STATISTICAL DESCRIPTION OF DATA

What Is Frequent Pattern Analysis?

Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set

First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining

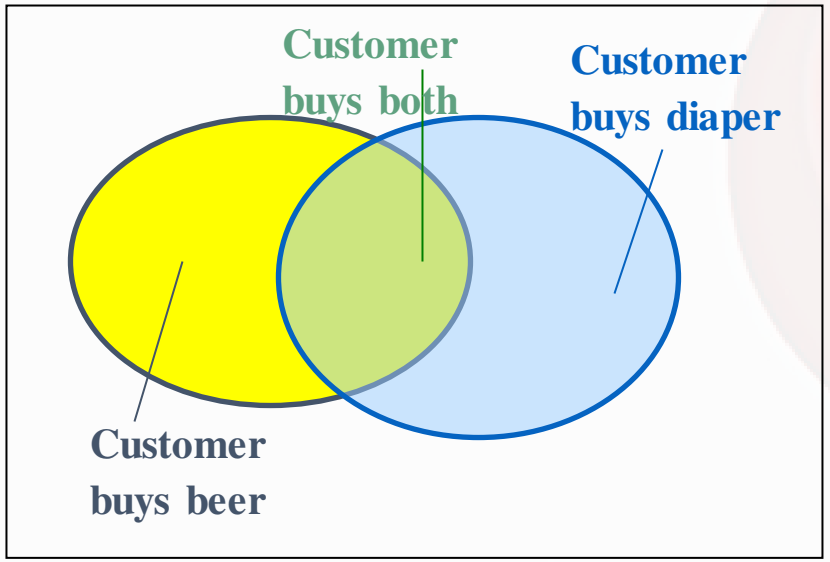
Motivation: Finding inherent regularities in data

- What products were often purchased together?— Beer and diapers?!
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we automatically classify web documents?

Applications

- Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold



Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Association Rules

- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, *s*, probability that a transaction contains $X \cup Y$
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

Let minsup = 50%, minconf = 50%

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

Closed Patterns and Max-Patterns

A long pattern contains a combinatorial number of sub-patterns, e.g., $\{a_1, \ldots, a_{100}\}$ contains $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{1}{1}\binom{0}{0}\binom{0}{0} = 2^{100} - 1 = 1.27*10^{30}$ sub-patterns!

Solution: *Mine closed patterns and max-patterns instead*

An itemset X is closed if X is *frequent* and there exists *no super-pattern* Y ↄ X, *with the same support* as X (proposed by Pasquier, et al. @ ICDT'99)

An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ↄ X (proposed by Bayardo @ SIGMOD'98)

Closed pattern is a lossless compression of freq. patterns

Reducing the # of patterns and rules

Closed Patterns and Max-Patterns

Exercise. DB = {<a_1, …, a_{100}>, < a_1, …, a_{50}>}

Min_sup = 1.

What is the set of closed itemset?

<a_1, …, a_{100}>: 1

< a_1, …, a_{50}>: 2

What is the set of max-pattern?

<a_1, …, a_{100}>: 1

What is the set of all patterns?

!!

References: Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791

Thank You