Data Preprocessing

——Discretization and Concept Hierarchy Generation——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

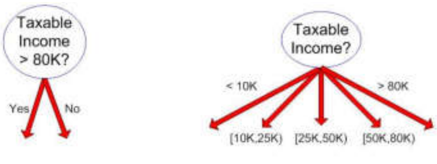
xuhua@tsinghua.edu.cn

Data Preprocessing

- About data
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Discretization and Concept hierarchy

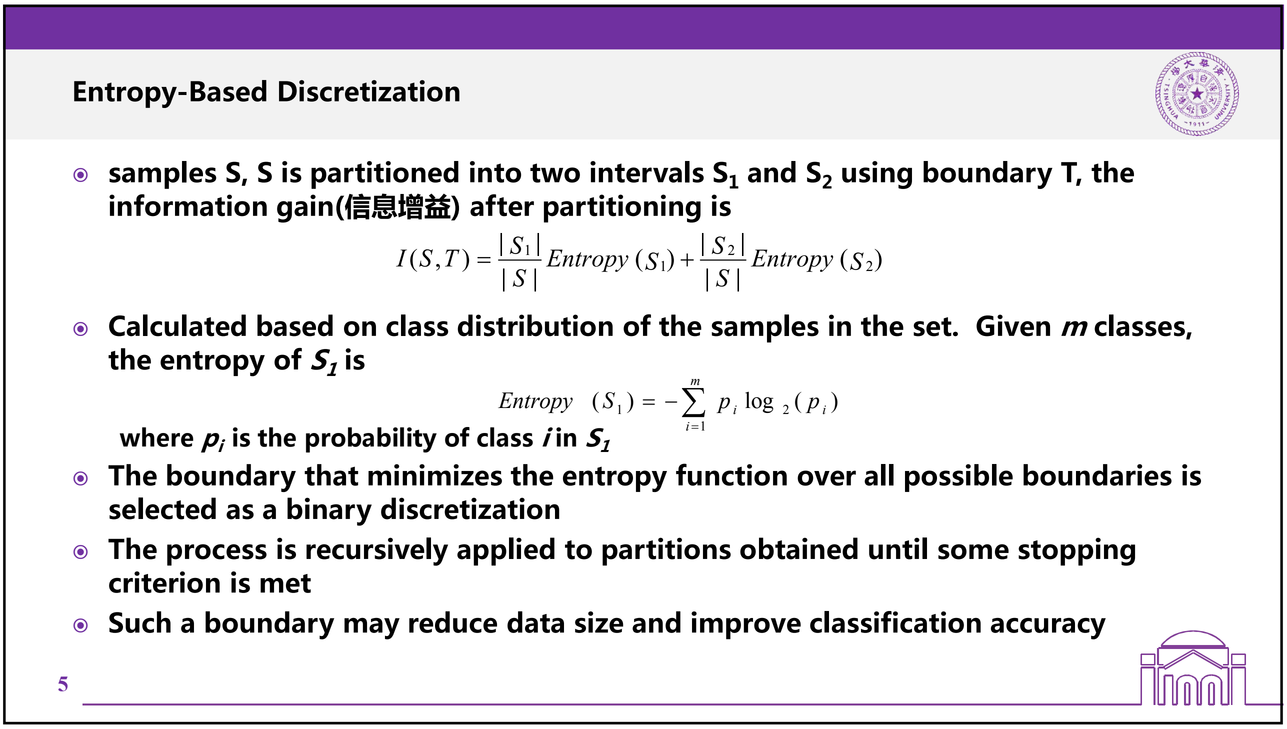
- Discretization
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values, e.g. salary, price, age
- Concept hierarchies
  - reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior) place-street-city-country



(i) Binary split (ii) Multi-way split

Discretization and Concept Hierarchy Generation for Numeric Data

- Binning (see sections before)
- Histogram analysis (see sections before)
- Clustering analysis (see sections before)
- Entropy-based discretization
- Segmentation by natural partitioning



Entropy-Based Discretization

- **samples S, S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain(信息增益) after partitioning is**

$$I(S,T) = \frac{|S_1|}{|S|} Entropy\,(S_1) + \frac{|S_2|}{|S|} Entropy\,(S_2)$$

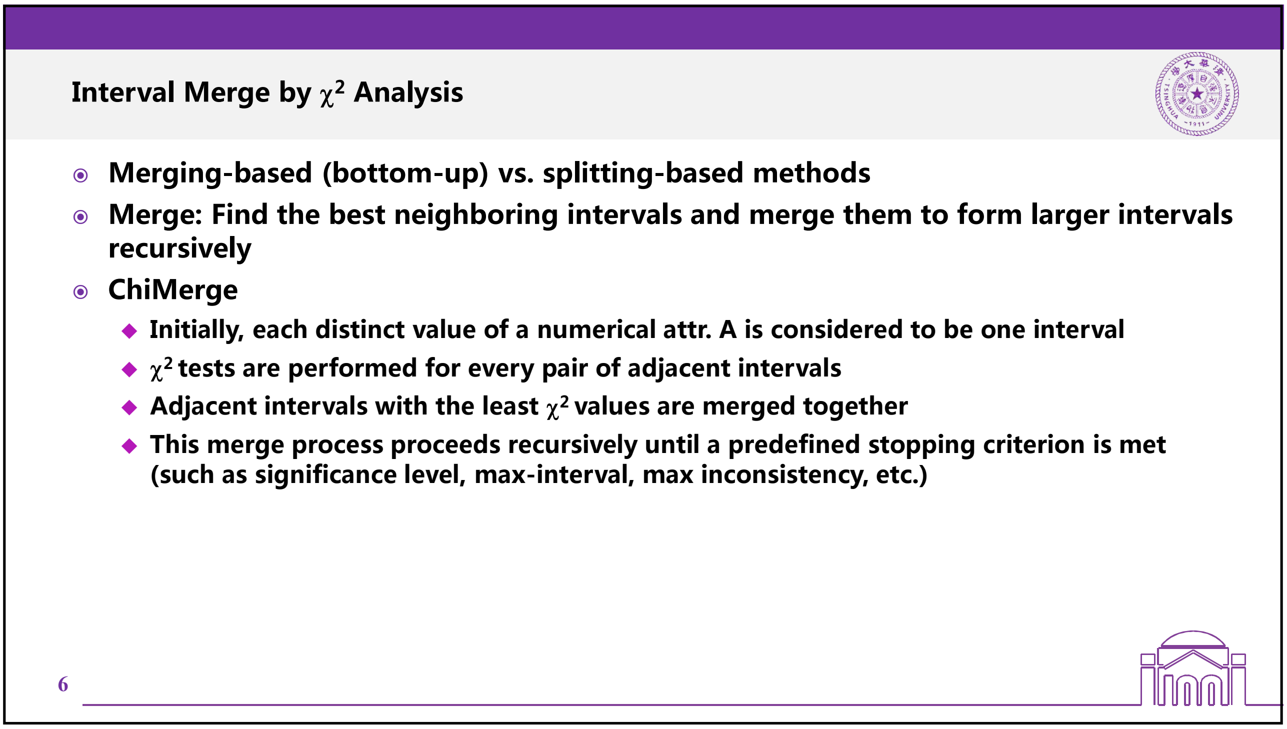
- **Calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of S_1 is**

$$Entropy\ (S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability of class *i* in S_1

- **The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization**
- **The process is recursively applied to partitions obtained until some stopping criterion is met**
- **Such a boundary may reduce data size and improve classification accuracy**

5



Interval Merge by χ^2 Analysis

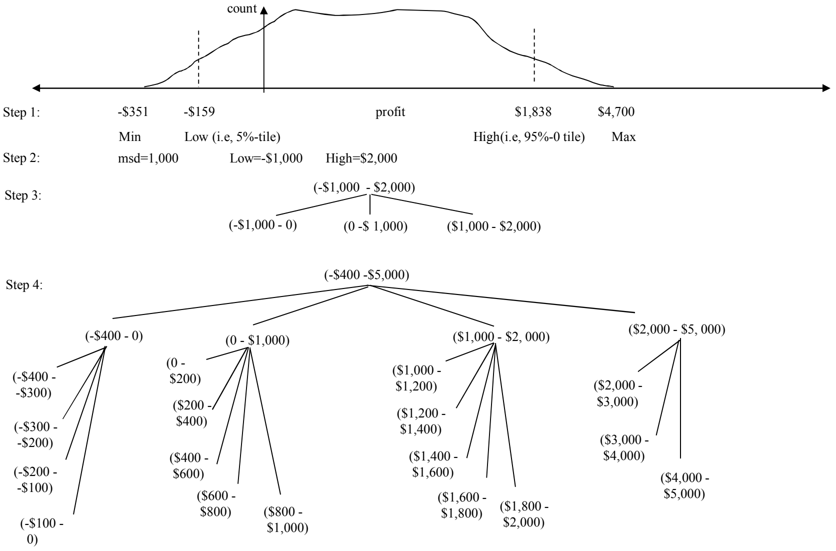
- **Merging-based (bottom-up) vs. splitting-based methods**
- **Merge: Find the best neighboring intervals and merge them to form larger intervals recursively**
- **ChiMerge**
  - **Initially, each distinct value of a numerical attr. A is considered to be one interval**
  - **χ^2 tests are performed for every pair of adjacent intervals**
  - **Adjacent intervals with the least χ^2 values are merged together**
  - **This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)**

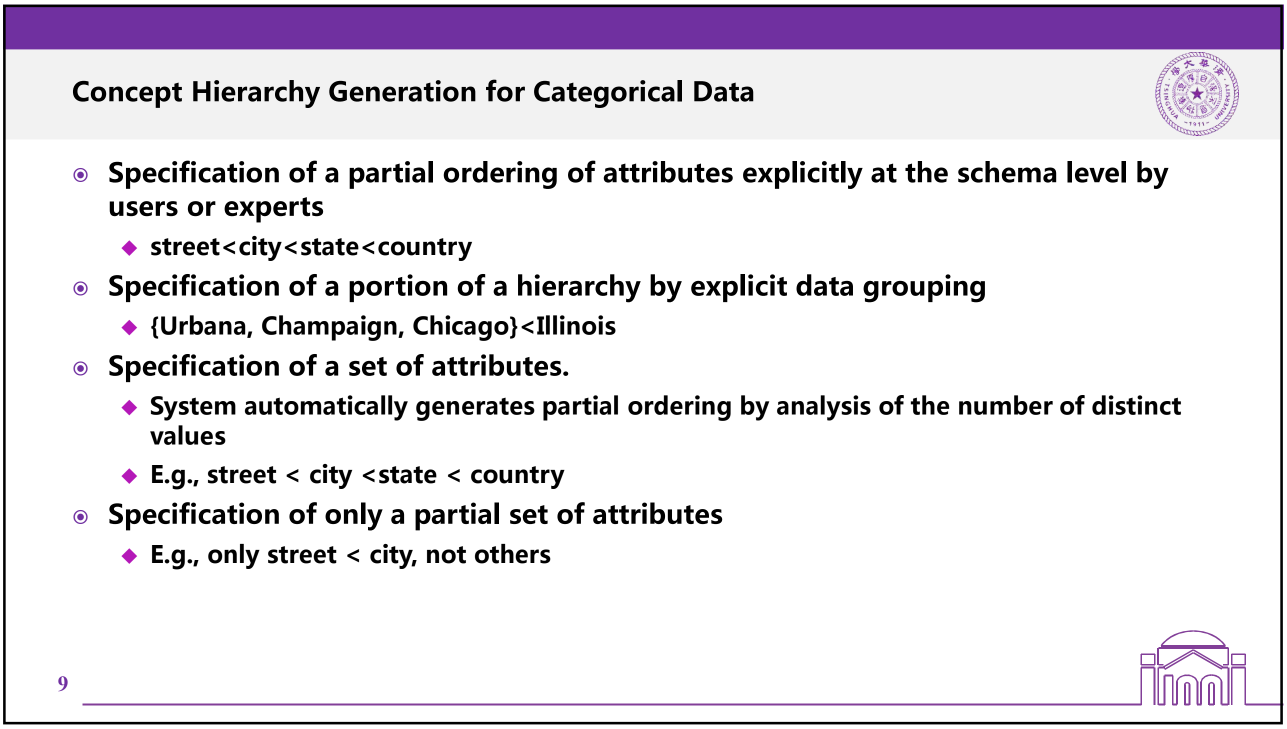
6

Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit(最高有效位), partition the range into 3 equi-width intervals
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Example of 3-4-5 Rule

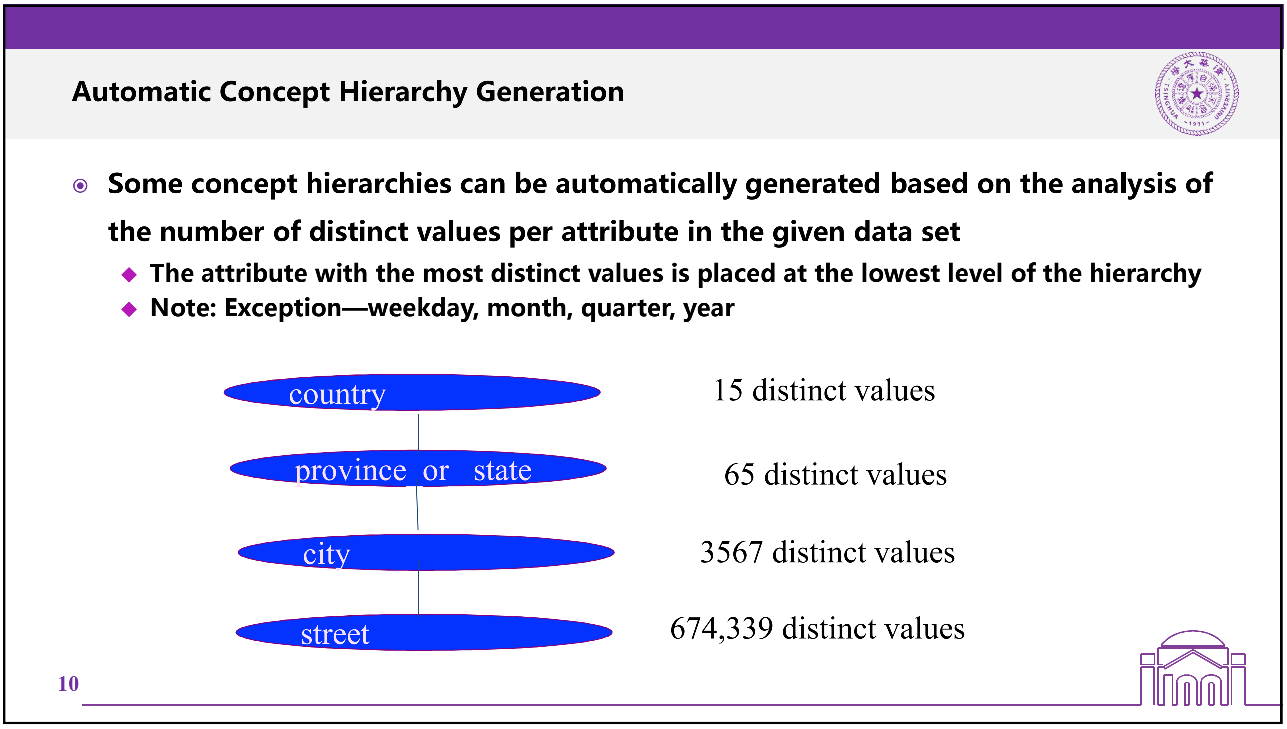




Concept Hierarchy Generation for Categorical Data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
  - street<city<state<country
- Specification of a portion of a hierarchy by explicit data grouping
  - {Urbana, Champaign, Chicago}<Illinois
- Specification of a set of attributes.
  - System automatically generates partial ordering by analysis of the number of distinct values
  - E.g., street < city <state < country
- Specification of only a partial set of attributes
  - E.g., only street < city, not others

9



Automatic Concept Hierarchy Generation

- Some concept hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the given data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Note: Exception—weekday, month, quarter, year

country	15 distinct values
province or state	65 distinct values
city	3567 distinct values
street	674,339 distinct values

10

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Discriptive data summarization is needed for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot of methods have been developed but data preprocessing still an active area of research

11

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. Communications of ACM, 39:86-95, 1996
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995

12

Thanks !

13