Data Mining

2.7 Data Discretization and Concept Hierarchy Generation

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

- Discretization and Concept Hierarchy Generation for Numerical Data
  - Binning
  - Entropy-Based Discretization
  - Interval Merging by $\chi 2$ Analysis
  - Clustering Analysis
- Concept Hierarchy Generation for Categorical Data
- References

Discretization and Concept Hierarchy Generation for Numerical Data

Data Discretization

- Data Discretization:
  - Dividing the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce the number of values for a given continuous attribute
  - Some classification algorithms only accept categorical attributes.
  - This leads to a concise, easy-to-use, knowledge-level representation of mining results.

Data Discretization

- Discretization techniques can be categorized based on whether it uses class information, as:
  - Supervised discretization
    - the discretization process uses class information
  - Unsupervised discretization
    - the discretization process does not use class information

Data Discretization

- Discretization techniques can be categorized based on which direction it proceeds, as:
  - Top-down
    - If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals
  - Bottom-up
    - starts by considering all of the continuous values as potential split-points,
    - removes some by merging neighborhood values to form intervals, and
    - then recursively applies this process to the resulting intervals.

Data Discretization

- Typical methods:
  - Binning
  - Entropy-based discretization
  - Interval merging by $\chi 2$ Analysis
  - Clustering analysis

- All the methods can be applied recursively
- Each method assumes that the values to be discretized are sorted in ascending order.

Binning

Binning

- The sorted values are distributed into a number of buckets, or bins, and then replacing each bin value by the bin mean or median
- Binning is:
  - a top-down splitting technique based on a specified number of bins.
  - an unsupervised discretization technique, because it does not use class information
- Binning methods:
  - Equal-width (distance) partitioning
  - Equal-depth (frequency) partitioning

Equal-width (distance) partitioning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

Equal-width (distance) partitioning

- Sorted data for price (in dollars):
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
- W = (B –A)/N = (34 – 4) / 3 = 10
  - Bin 1: 4-14, Bin2: 15-24, Bin 3: 25-34
- Equal-width (distance) partitioning:
  - Bin 1: 4, 8
  - Bin 2: 15, 21, 21, 24
  - Bin 3: 25, 28, 34

Equal-depth (frequency) partitioning

- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Equal-depth (frequency) partitioning

- Sorted data for price (in dollars):
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
- Equal-depth (frequency) partitioning:
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34

Entropy-Based Discretization

Entropy-Based Discretization

- Entropy-based discretization is a **supervised**, **top-down splitting** technique.
- It explores class distribution information in its calculation and determination of split-points
- Let D consist of data instances defined by a set of attributes and a class-label attribute.
- The class-label attribute provides the class information per instance.

Entropy-Based Discretization

- The basic method for entropy-based discretization of an attribute A within the set is as follows:

1) Each value of A can be considered as a potential interval boundary or split-point (denoted split point) to partition the range of A.
   - That is, a split-point for A can partition the instances in D into two subsets satisfying the conditions A $\leq$ ***split_point*** and A **>** ***split_point***, respectively,
   - thereby creating a binary discretization.

Entropy-Based Discretization

2) the information gain after partitioning is

$$Info_A(D) = \frac{|D_1|}{|D|}Entropy(D_1) + \frac{|D_2|}{|D|}Entropy(D_2)$$

- where D_1 and D_2 correspond to the instances in D
- |D| is the number of instances in D, and so on.
- The entropy function for a given set is calculated based on the class distribution of the tuples in the set.
- For example, given m classes, C_1, C_2, …, C_m, the entropy of D_1 is:

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Entropy-Based Discretization

– where p_i is the probability of class C_i in D_1, determined by dividing the number of tuples of class C_i in D_1 by $|D_1|$, the total number of tuples in D_1.

– Therefore, when selecting a split-point for attribute A, we want to pick the attribute value that gives the minimum expected information requirement (i.e., $\min(Info_A(D))$).

3) The process of determining a split-point is recursively applied to each partition obtained, until some stopping criterion is met, such as:

– when the minimum information requirement on all candidate split-points is less than a small threshold, e,

– or when the number of intervals is greater than a threshold, *max_interval.*

Entropy-Based Discretization

- The interval boundaries (split-points) are defined may help improve classification accuracy
- The entropy and information gain measures described here are also used for decision tree induction.

Interval Merge by χ^2 Analysis

Interval Merge by χ^2 Analysis

- ChiMerge:
  - It is a bottom-up method
  - Find the best neighboring intervals and merge them to form larger intervals recursively
  - The method is supervised in that it uses class information.
  - The basic notion is that for accurate discretization, the relative class frequencies should be fairly consistent within an interval.
  - Therefore, if two adjacent intervals have a very similar distribution of classes, then the intervals can be merged. Otherwise, they should remain separate.
  - ChiMerge treats intervals as discrete categories

Interval Merge by χ^2 Analysis

- The ChiMerge method:
  - Initially, each distinct value of a numerical attribute 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, since low χ^2 values for a pair indicate similar class distributions
  - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

Cluster Analysis

Cluster Analysis

- Cluster analysis is a popular data discretization method.
- A clustering algorithm can be applied to discretize a numerical attribute, A, by partitioning the values of A into clusters or groups.
- Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.

Cluster Analysis

- Clustering can be used to generate a concept hierarchy for A by following either a top-down splitting strategy or a bottom-up merging strategy, where each cluster forms a node of the concept hierarchy.
- In the former, each initial cluster or partition may be further decomposed into several subclusters, forming a lower level of the hierarchy.
- In the latter, clusters are formed by repeatedly grouping neighboring clusters in order to form higher-level concepts.

Concept Hierarchy Generation for Categorical Data

Concept Hierarchy Generation for Categorical Data

- Generalization is the generation of concept hierarchies for categorical data
- Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values.
- Examples include
  - geographic location,
  - job category, and
  - itemtype.

Concept Hierarchy Generation for Categorical Data

- There are several methods for the generation of concept hierarchies for categorical data:
  - Specification of a partial ordering of attributes explicitly at the schema level by users or experts
  - Specification of a portion of a hierarchy by explicit data grouping
  - Specification of a set of attributes, but not of their partial ordering

Concept Hierarchy Generation for Categorical Data

- **Specification of a partial ordering of attributes explicitly at the schema level by users or experts**
  - Example: a relational database or a dimension location of a data warehouse may contain the following group of attributes: street, city, province or state, and country.
  - A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
  - A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as:
    - **street < city < province or state < country**

Concept Hierarchy Generation for Categorical Data

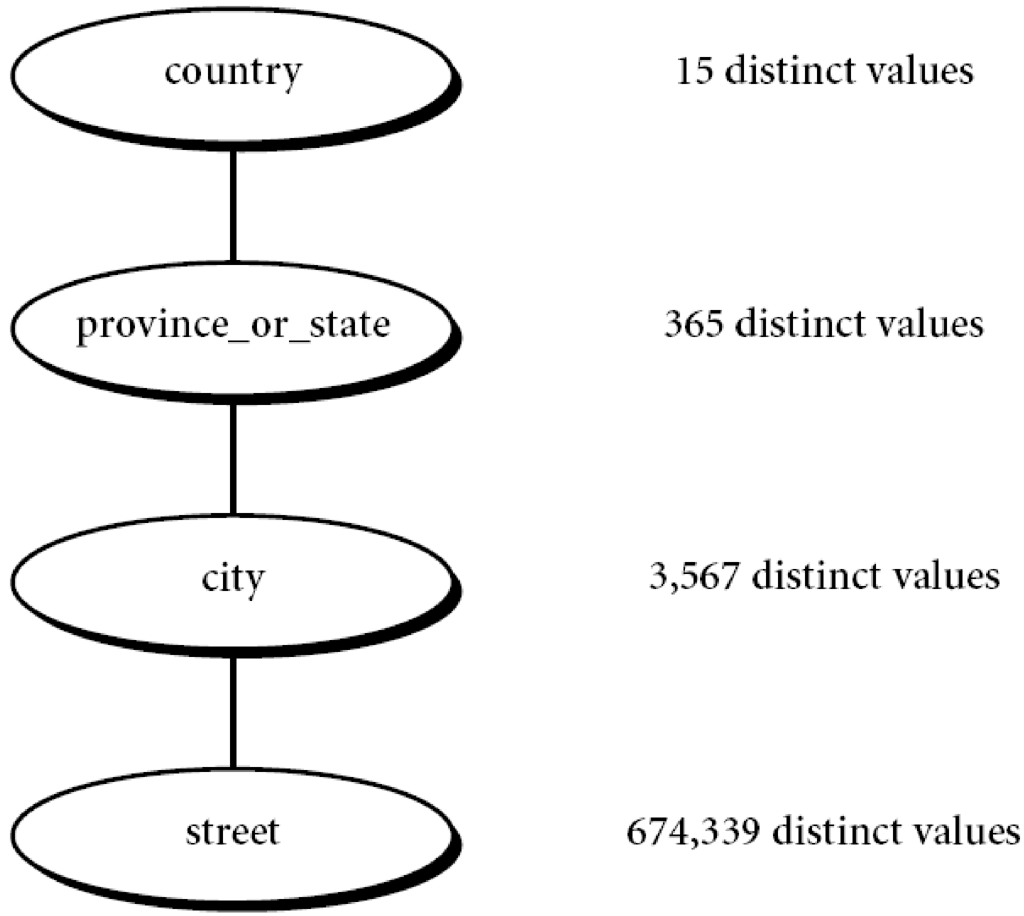
- **Specification of a portion of a hierarchy by explicit data grouping**
  - we can easily specify explicit groupings for a small portion of intermediate-level data.
  - For example, after specifying that province and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as:
    - **{Urbana, Champaign, Chicago} < Illinois**

Concept Hierarchy Generation for Categorical Data

- **Specification of a set of attributes, but not of their partial ordering**
  - A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
  - The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.
  - Example: Suppose a user selects a set of location-oriented attributes, *street, country, province_or_state,* and *city,* from the *AllElectronics* database, but does not specify the hierarchical ordering among the attributes.

Concept Hierarchy Generation for Categorical Data

- Automatic generation of a schema concept hierarchy based on the number of distinct attribute values.
- The attribute with the most distinct values is placed at the lowest level of the hierarchy
- Exceptions, e.g., weekday, month, quarter, year



References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end

- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)