Data Mining (CS 451)

JAN-MAY' 18

PRAGYA VERMA

Recap

- What is Data Mining?
- Why Data Mining?
- What Kinds of Data can be Mined?
- Overview of Different Data Mining Tasks

CONTENTS

Types of Data

- Most data falls into one of the two groups – numerical or categorical
- Numerical Data: Data that can be quantified is known as numerical data. For example: A person’s height, weight, blood pressure.
- Numerical data can be further broken down into two types: discrete and continuous.
- Discrete data represent items that can be counted; they take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity (making it countably infinite).

Types of Data (Contd.)

- Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line.
- Categorical data represent characteristics such as a person’s gender, marital status, hometown, or the types of movies they like.
- Categorical data can take on numerical values (such as “1” indicating male and “2” indicating female), but those numbers don’t have mathematical meaning. You couldn’t add them together, for example.
- This type of data is also known as qualitative data, or Yes/No data.

Types of Data (Contd.)

- Ordinal data mixes numerical and categorical data. The data fall into categories, but the numbers placed on the categories have meaning.
- For example, rating a restaurant on a scale from 0 (lowest) to 4 (highest) stars gives ordinal data. Ordinal data are often treated as categorical, where the groups are ordered when graphs and charts are made.
- However, unlike categorical data, the numbers do have mathematical meaning.
- For example, if you survey 100 people and ask them to rate a restaurant on a scale from 0 to 4, taking the average of the 100 responses will have meaning.

Importance of Data Preprocessing

- Today's real world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources.
- Low quality data will lead to low quality mining results
- Data preprocessing techniques when applied before mining can substantially improve the overall qualities of the patterns mined and/or the time required for the actual mining.

Data Quality: Why Preprocess the Data?

- Data is said to be of good quality if they satisfy the requirements of the intended use
- There are many factors comprising data quality:

1. Accuracy
2. Completeness
3. Believability
4. Consistency
5. Timeliness
6. Interpretability

Data Quality: Why Preprocess the Data?

- Incomplete: Lacking attribute values or certain attributes of interest or containing only aggregate data
- Inaccurate or noisy: Containing errors, or values that deviate from expected
- Inconsistent: Containing discrepancies.

Data Quality: Why Preprocess the Data?

- Timeliness: Data needs to be updated in a timely manner, otherwise it may have a negative impact on the data quality
- Believability: Reflects how much the data is trusted by the users
- Interpretability: Reflects how easily the data is understood

Data Quality: Why Preprocess the Data?

- Reasons for poor data quality can be:

1. Data collection instruments may be faulty
2. Human or computer errors occurring at data entry
3. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information. This is also known as disguised missing data

- Data quality depends on the intended use of data
- Two different users may have a very different assessment of the quality of a given database.

Data Cleaning

- Real-world data tend to be incomplete, noisy and inconsistent.
- If users believe that the data is dirty, they are unlikely to trust the results of any data mining that has been applied
- Dirty data can cause confusion for the mining procedure, resulting in unreliable output.
- Quality decisions can only be made using quality data

Data Cleaning (Contd.)

- Data cleaning works to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- It is also known as data cleansing.

Data Cleaning – Missing Values

- Imagine that you need to analyze customer data. You may note that many tuples have no recorded value for several attributes such as customer income.
- How can you go about filling in the missing values for this attribute?
- One of the following techniques can be used to fill in the missing value.

Data Cleaning – Missing Values

1. Ignore the tuple:
   - This method is not effective, unless the tuple contains several attributes with missing values
   - Poor, when the percentage of missing value per attribute varies considerably
   - By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple
2. Fill in the missing values manually:
   - Time Consuming
   - May not be feasible given a large dataset with many missing values

Data Cleaning – Missing Values

3. Use of global constant to fill in the missing value:
   - Replace all missing attribute values by the same constant such as a label like "unknown".
4. Use a measure of central tendency for the attribute:
   - Central tendency indicates the "middle" value of a data distribution
   - For a normal data distribution, mean can be used, while skewed data distribution should employ median.
5. Use the most probable value to fill in the missing value:
   - This may be determined with regression, inference-based tools, or decision tree
   - For example, using other customer attributes in the data set, you may construct a decision tree to predict the missing values for income.

Data Cleaning – Noisy Data

- Noise is a random error or variance in a measured variable.
- Outliers may represent noise.
- Given a numeric attribute such as, say, price, how can we “smooth” out the data to remove noise?

Data Cleaning – Noisy Data

- The following smoothing techniques can be used:

1. Binning:
   - Binning methods smooth a sorted data value by consulting its “neighborhood”, that is, the values around it.
   - The sorted values are distributed into a number of buckets or bins.
   - In **smoothing by bin means**, each value in a bin is replaced by the mean value of the bin.
   - Similarly, **smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.
   - In **smoothing by bin boundaries**, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Next Class

- Data Integration