Data Pipeline

Introduction to Data Mining

What is data mining

- Knowledge discovery from data
- Extraction of interesting patterns or knowledge from data

Data Mining 4 view

- Data
    - 5 Vs of data
        - Value
        - Volume
        - Variety
        - Velocity
        - Veracity
    - Relational, transactional
    - Sequential, temporal
    - Spatial
    - Text, multimedia
    - Graph, network

- Application
    - Market analysis
    - Healthcare, medical research
    - Science and engineering
    - Security
    - Government, nonprofit

- Knowledge
    - Frequent pattern, association, correlation
    - Categorization
    - Anomaly, outliers
    - Changes over time

- Technique
    - Frequent pattern analysis
        - Itemset
        - Sequence
        - Structure
        - Association
        - Correlation

    - Classification, prediction
        - Pre-defined
        - Training data
        - Modeling
    - Clustering
        - No predefined classes
        - Intra-cluster similarity
        - Inter-cluster dissimilarity
    - Anomaly detection
    - Trend and evolution analysis

Data Mining Pipeline Structure

- Data Understanding
- Data Preprocessing
    - Potential issues with data
    - Preparing data for mining
    - Garbage in, garbage out
- Data Warehousing
    - Data Warehouse
    - Data cub and OLAP
    - Data warehouse structure
- Data Modeling
    - Frequent pattern analysis
    - Classification, prediction
    - Clustering
    - Anomaly detection
    - Trend and evolution analysis
- Pattern Evaluation
    - Evaluation metrics
        - Accuracy, error rate
        - False + / - rate
        - Efficiency, latency
    - Model selection

Major Issues in Data Mining

- Diverse data -> diverse knowledge
- Data quality
- Supervised vs unsupervised learning
- Performance evaluation
- Effectiveness vs efficiency
- Incremental, interactive mining
- Integration of domain knowledge
- Privacy-preserving mining

Data Ethics

- Data ownership
- Privacy, anonymity
- Data and model validity
- Data and model bias
- Interpretation, application, societal consequence

Data Understanding

Data Objects and Attributes

- Dataset
    - A collection of data objects
    - Each described by a number of attributes
- Attribute types
    - Categorical
    - Numeric
    - Discrete
    - Continuous

Data Statistics

- Distribution of each attribute's values
    - Categorial
        - Percent of each values
    - Numeric
        - Central Tendency
            - Mean, median, mode, midrange (max-min)/2
    - Dispersion
        - How much a distribution is stretched or squeezed
        - Range: max-min
        - Quartiles: Q1 25%, Q3 75%
        - IQR: Q3-Q1
        - Variance
        - Standard deviation

Data Visualization

- Boxplot
    - Box
        - Q1, Q2, Q3, IQR
    - Whiskers
        - Min, max, 1.5 x IGR
    - Outliers

- Histogram
    - Bars of different height
    - X: subrange, bins
    - Y: frequency, bar height
- Quantile plot
    - Percent of points below a given values
    - X: percent
    - Y: Quantile
- Q-Q plot
    - Comparison of 2 quantiles
    - 45-degree ref line
- Scatter plot

Data Similarity

- Object matrix
    - n ob x p att
- Dissimilarity matrix
    - n ob x n ob
- Nominal Attributes
    - Similarity
        - s=1 if x=y otherwise s=0
    - Dissimilarity
        - d=0 if x=y otherwise d=1
- Binary Attributes
    - Symmetric
    - Asymmetric
- Ordinal Attributes
    - Map to their ranks
    - Map to 0,1
    - Dissimilarity between mapped values
- Numeric Object Dissimilarity
    - Usually measured by distance
    - Minkowski distance
- Cosign Similarity
    - Angular similarity of vectors
- Sequential Data, Time Series
    - Euclidean distance
    - Dynamic time warping
    - Minimum jump cost
- Mixed Attribute Types
    - Weighted sums across attributes
- Data Similarity / Dissimilarity
    - How to choose

        - Dense, continuous: Euclidean, Manhattan
        - Asymmetric: Ignore the null
        - Sparse: Cosine similarity, Jaccard similarity
        - Subset: Seasonal patterns, subgroups
        - Domain knowledge

Data Preprocessing

Data Quality

- Relevance
- Accessibility
- Interpretability
- Timeliness
- Accuracy
- Consistency
- Precision
- Granularity
- Completeness

Real World Issues

- Incomplete: missing values
- Noisy: imprecise, errors, outliers
- Inconsistent

Data Cleaning

- Incomplete:
    - Remove objects or attributes
    - Manually fill in missing values
    - Automated methods
        - Attribute mean, kNN, etc
- Noisy
    - Regression
        - Fill data with regression functions
    - Clustering
        - Group data
        - Remove outliers
- Inconsistent
    - Semantic bases checking
        - Metadata, attribute relationships
    - Data understanding
        - Statistical analysis

Data Integration

- Combines data from multiple sources
- Entity identification
    - Users, items
- Redundant data
    - Correlation

Correlation

- Numerical attribution
    - Correlation coefficient
- Nominal attribution
    - Chi-square

Data Transformation

- Normalization
    - Rescaling
        - Min-Max normalization
    - Mean normalization
    - Standardization
        - Z score normalization
- Discretization
    - Continuous: intervals
    - Split or merge
    - Supervised or unsupervised: class labels
        - Unsupervised
            - Binning and histogram
            - Cluster analysis
            - Intuitive partitioning
        - Supervised
            - Pre-determined class labels
            - Entropy-based interval splitting
                - Lower entropy-> purer class
            - X2 analysis-based interval merging
                - Lower X2 means class is independent
- Data Reduction
    - Dimensionality reduction
        - Attribute selection
            - Forward
            - Backward
            - Feature engineering
        - Principle Component Analysis
            - n dimensional data
        - Wavelet Transformation

        - Linear signal processing
    - Numerosity reduction: objects
        - Parametric
        - Non-parametric

Data Warehousing

Data Warehouse

- William H Inmon: A subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision-making process
- Separated from operational data
- Keey Characteristics
    - Subject oriented
    - Integrated
    - Time-variant
    - Nonvolatile

OLTP vs OLAP

- Online Transactional Processing OLTP
    - Transactional-oriented tasks: purchase, bank transaction
- Online Analytical Processing OLAP
    - Complete queries on historical data

Data Warehouse Model

- Fact vs dimension
- Star schema
    - 1 fact table, multiple dimension tables
- Snowflake schema
    - 1 fact table, multiple levels of dimension tables
- Fact constellation schema
    - Multiple fact tables, shared dimension tables

Data Cube

- Multidimensional data model
- Roll up: aggregation
- Drill down: reverse roll up
- Pivot: rotate
- Slicing: select along a single dimension
- Dicing: select along multiple dimensions
- Materialization
    - Full

        - Multiway Array Aggregation
            - Bottom-up computation
            - Simultaneously aggregate along multiple dimensions
    - No
        - ROLAP
    - Partial
        - Iceberg Cube
        - Star Cubing
            - Bottom-up computation
            - Top-down expansion
        - Bottom-up Computation BUC
            - Top-down
            - Iceberg pruning

Staging ETL

- Extract
- Transform
- Load