
Lecture 13: Data Mining

Dragan Mirkovic
Department of Computer Science
University of Houston

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Announcements

- Today:
 - A short overview of data mining
 - Chapter 26 in Ramakrishnan & Gehrke
- Thursday:
 - Review for the final exam
 - Instructor evaluations
 - The final exam is scheduled for Tuesday, Dec 14th, 11 am-2 pm.

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Introduction

- **Data mining:** Finding useful trends and patterns in large datasets to guide future decisions
 - Set of methods and tools for analysis and modeling of very large amounts of data
 - Scalability is very important
 - Related scientific areas:
 - Statistics:
 - Exploratory data analysis
 - Artificial intelligence:
 - Knowledge discovery and machine learning

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

An Abundance of Data

- Supermarket scanners, POS data
- Preferred customer cards
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Demographic data
- Sensor networks
- Cameras
- Web server logs
- Customer web site trails
- High-throughput scientific instruments
- Increase in computational power
- Moore's Law:
In 1965, Intel Corporation cofounder Gordon Moore predicted that the density of transistors in an integrated circuit would double every year. (Later changed to reflect 18 months progress.)
- Result: **Data mining**

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Why Use Data Mining Today?

- Competitive pressure!
 - “The secret of success is to know something that nobody else knows.” - Aristotle Onassis
- Competition on service, not only on price (Banks, phone companies, hotel chains, rental car companies)
- Personalization, CRM
- The real-time enterprise
- “Systemic listening”
- Security, homeland defense

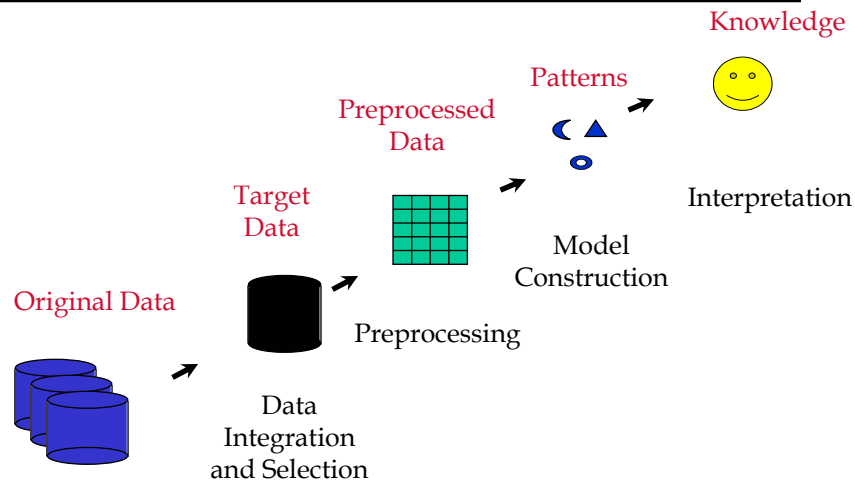
D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Knowledge Discovery Process

- The steps in KDD process:
 - Data preprocessing
 - Data selection: Identify target datasets and relevant fields
 - Data cleaning
 - Remove noise and outliers
 - Data transformation
 - Create common units
 - Generate new fields
 - Data mining model construction
 - Extract useful patterns by using data mining algorithms
 - Model evaluation
 - Presentation of models through visualization

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Knowledge Discovery Process



D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Example Application: Sports

- IBM Advanced Scout analyzes NBA game statistics
 - A large number of data generated for each game, player, teams, etc. (see <http://www.nba.com>)
 - Points per game, rebounds per game, field goal percentage, 3 point field goal percentage, free throw percentage, assists per game, assists per turnover, steals per game, blocks per game, turnovers per game, fouls per game, ...
- Used by NBA coaching staffs to discover interesting patterns in basketball game data.



D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Advanced Scout - Article

U.S. News: Basketball's New High-Tech Guru

by Scott McMurray | Dec 11 '95

The New York Knicks coaching staff filed glumly out of Madison Square Garden after the Knicks barely eked out a win over the struggling Vancouver Grizzlies expansion franchise last month. The next morning, at the team's Westchester County practice center north of Manhattan, a scout pinpointed a major reason the victory had been such a close call. Two of the Knicks players, Patrick Ewing and Anthony Mason, hadn't made nearly as many baskets as usual from the "post up" position near the free-throw line. Ewing and Mason were each stopping several feet farther away from the basket than they were supposed to, making it that much harder for them to hit their shots.

Who discovered the glitch in the Knicks' game? Advanced Scout, IBM's state-of-the-art computer software, which is currently being offered free of charge to all National Basketball Association teams. By crunching reams of statistics at warp speed, the innovative new program helps coaches analyze their teams' performance. ...

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Example Application: Sky Survey

- Input data: 3 TB of image data with 2 billion sky objects, took more than six years to complete
- Goal: Generate a catalog with all objects and their type
- Method: Use decision trees as data mining model
- Results:
 - 94% accuracy in predicting sky object classes
 - Increased number of faint objects classified by 300%
 - Helped team of astronomers to discover 16 new high red-shift quasars in one order of magnitude less observation time
- News:
 - Sloan Digital Sky Survey Finds Mysterious New Milky Way Companion (10/20/2004)



D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

SQL/MM

- Data Mining extension of the SQL
- Supports four kinds of data mining models:
 - Frequent itemsets and association rules
 - Clusters of records
 - Regression trees
 - Classification trees
- Support for new data types required by the data mining models
- Models can be exported in a standard XML format called PMML
 - Predictive Model Markup Language

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

What is a Data Mining Model?

- A data mining model is a description of a specific aspect of a dataset. It produces output values for an assigned set of input values.
- Examples:
 - Linear regression model
 - Classification model
 - Clustering

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Data Mining Models (Contd.)

A data mining model can be described at two levels:

- Functional level:
 - Describes model in terms of its intended usage.
Examples: Classification, clustering
- Representational level:
 - Specific representation of a model.
Example: Log-linear model, classification tree, nearest neighbor method.
- Black-box models versus transparent models

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Data Mining: Types of Data

- Relational data and transactional data
- Spatial and temporal data, spatio-temporal observations
- Time-series data
- Text
- Images, video
- Mixtures of data
- Sequence data
- Features from processing other data sources

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Types of Variables

- *Numerical*: Domain is ordered and can be represented on the real line (e.g., age, income)
- *Nominal* or *categorical*: Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- *Ordinal*: Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Data Mining Techniques

- Supervised learning
 - Classification and regression
- Unsupervised learning
 - Clustering
- Dependency modeling
 - Associations, summarization, causality
- Outlier and deviation detection
- Trend analysis and change detection

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Counting Co-Occurrences

- Motivation:
 - Market basket analysis
 - Collection of items purchased in a single customer transaction
 - Examples:
 - Almost all internet retailers like Amazon or eBay will tempt you with the statistics.
 - Improvement of the layout of goods in a store or a catalog

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Frequent Itemsets

- **Itemset**: a set of items
 - Example: {pen, ink}
- The **support** of an itemset: a fraction of transactions in the database that contain all the items in the itemset
 - Example: {pen, ink} has 75% support in Purchases
- The a priori property: every subset of a frequent itemset is also a frequent itemset
- Iterations:
 - One item itemsets, two items itemsets, ...

The Purchases Relation

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Market Basket Analysis

- Co-occurrences
 - 80% of all customers purchase items X, Y and Z together.
- Association rules
 - 60% of all customers who purchase X and Y also buy Z.
- Sequential patterns
 - 60% of customers who first buy X also purchase Y within three weeks.

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Confidence and Support

We prune the set of all possible association rules using two interestingness measures:

- **Confidence** of a rule:
 - $X \rightarrow Y$ has confidence c if $P(Y|X) = c$
- **Support** of a rule:
 - $X \rightarrow Y$ has support s if $P(XY) = s$

We can also define

- **Support** of an itemset (a cocurrence) XY :
 - XY has support s if $P(XY) = s$

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Example

Examples:

- {Pen} => {Milk}
Support: 75%
Confidence: 75%
- {Ink} => {Pen}
Support: 100%
Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Market Basket Analysis: Applications

- Sample Applications
 - Direct marketing
 - Fraud detection (medical insurance, credit cards)
 - Floor/shelf planning
 - Web site layout
 - Cross-selling

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Clustering

- **Input:**
 - A data set of N records each given as a d-dimensional data feature vector.
- **Output:**
 - Determine a natural, useful “partitioning” of the data set into a number of (k) clusters and noise such that we have:
 - High similarity of records within each cluster (intra-cluster similarity)
 - Low similarity of records between clusters (inter-cluster similarity)
- **Distance function:**
 - measures similarity between records
- **Center** of a collection of records, r_1, \dots, r_n $C = \frac{1}{n} \sum_{i=1}^n r_i$
- **Radius** of a collection of records, r_1, \dots, r_n $R = \sqrt{\frac{1}{n} \sum_{i=1}^n d(r_i, C)}$

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Types of Clustering

- **Hard Clustering:**
 - Each object is in one and only one cluster
- **Soft Clustering:**
 - Each object has a probability of being in each cluster

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Clustering Algorithms

- Partitioning-based clustering
 - K-means clustering
 - K-medoids clustering
 - EM (expectation maximization) clustering
- Hierarchical clustering
 - Divisive clustering (top down)
 - Agglomerative clustering (bottom up)
- Density-Based Methods
 - Regions of dense points separated by sparser regions of relatively low density

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

K-Means Clustering Algorithm

- Initialize k cluster centers
 - Do**
 - Assignment step:** Assign each data point to its closest cluster center
 - Re-estimation step:** Re-compute cluster centers
 - While** (there are still changes in the cluster centers)
- Advantages:
 - Good for exploratory data analysis
 - Works well for low-dimensional data
 - Reasonably scalable
- Disadvantages
 - Hard to choose k
 - Often clusters are non-spherical

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

K-Medoids

- Similar to K-Means, but for categorical data or data in a non-vector space.
- Since we cannot compute the cluster center (think text data), we take the “most representative” data point in the cluster.
- This data point is called the medoid (the object that “lies in the center”).

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Agglomerative Clustering

- Algorithm:
 - Put each item in its own cluster (all singletons)
 - Find all pairwise distances between clusters
 - Merge the two closest clusters
 - Repeat until everything is in one cluster
- Observations:
 - Results in a hierarchical clustering
 - Yields a clustering for each possible number of clusters
 - Greedy clustering: Result is not “optimal” for any cluster size

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Density-Based Clustering

- A cluster is defined as a connected dense component.
- Density is defined in terms of number of neighbors of a point.
- We can find clusters of arbitrary shape



D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004

Density-Based Clustering

- Advantages:
 - Finds clusters of arbitrary shapes
- Disadvantages:
 - Targets low dimensional spatial data
 - Hard to visualize for >2-dimensional data
 - Needs clever index to be scalable
 - How do we set the magic parameters?

D. Mirkovic, COSC 3480: Design of File and Database Systems, Fall 2004