

Knowledge Discovery and Data Mining

Unit # 1

Course Outlines

- Classification Techniques
 - Classification/Decision Trees
 - Naïve Bayes
 - Neural Networks
- Clustering
 - Partitioning Methods
 - Hierarchical Methods
- Patterns and Association Mining
- Time Series and Sequence Data
- Data Visualization
- Anomaly Detection
- Papers/Case Studies/Applications Reading

Software/Data Repository

- Weka
- KNIME
- R
- SQL Server (Business Intelligence)
- UCI Machine Learning Repository
 - <http://www.ics.uci.edu/~mlearn/>

Useful Information

- Course Wiki
 - <http://cse652fall2011.wikispaces.com>
- Text/Reference Books
 - Introduction to Data Mining by Tan, Steinbach and Kumar (2006)
 - Data Mining Concepts and Techniques by Han and Kamber (2006)
 - Data Mining: Practical Machine Learning Tools and Techniques by Witten and Frank (2005)

Acknowledgement

- Several Slides in this presentation are taken from course slides provided by
 - Han and Kimber (Data Mining Concepts and Techniques) and
 - Tan, Steinbach and Kumar (Introduction to Data Mining)

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

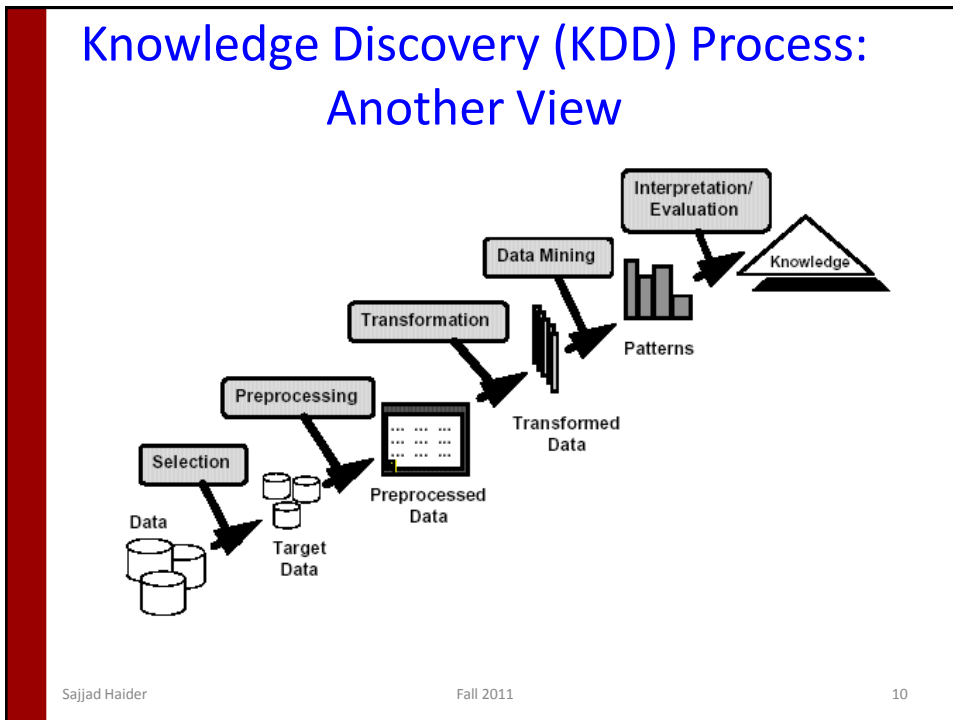
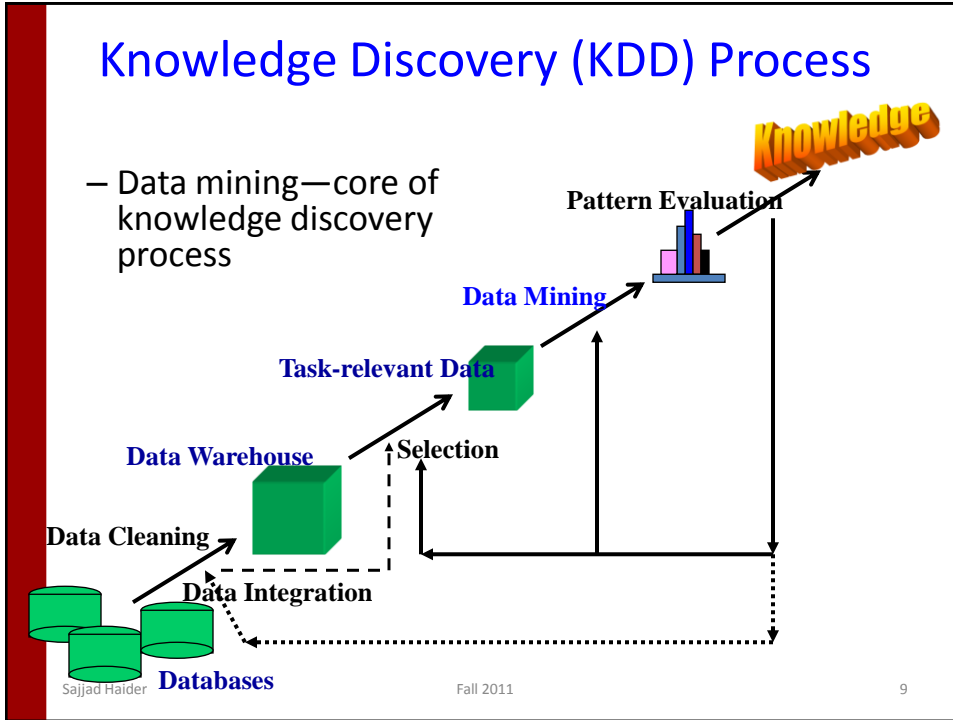


What Is Data Mining?

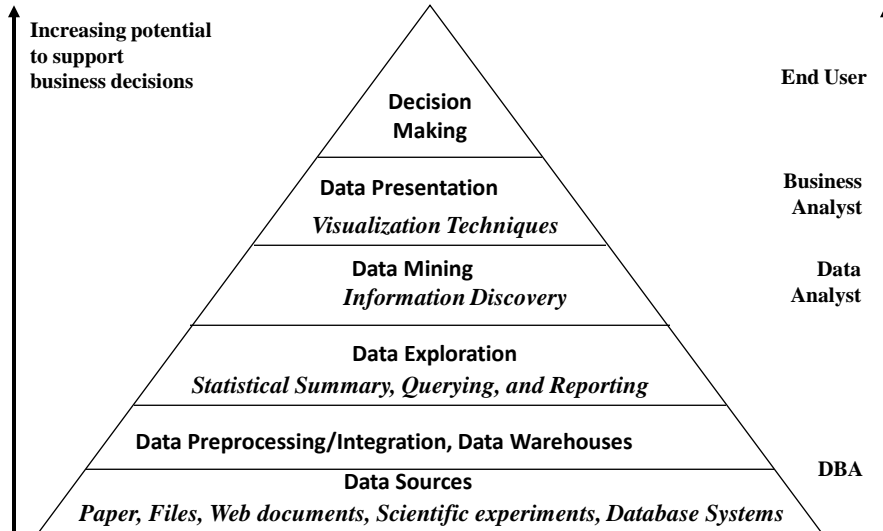


- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems

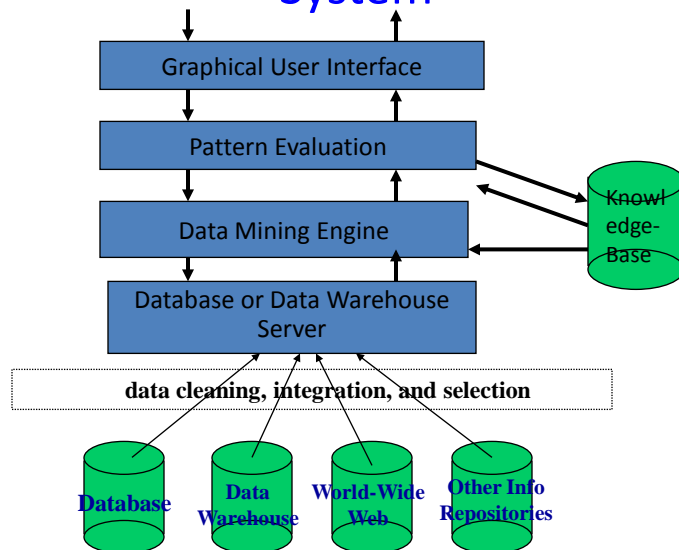




Data Mining and Business Intelligence



Architecture: Typical Data Mining System

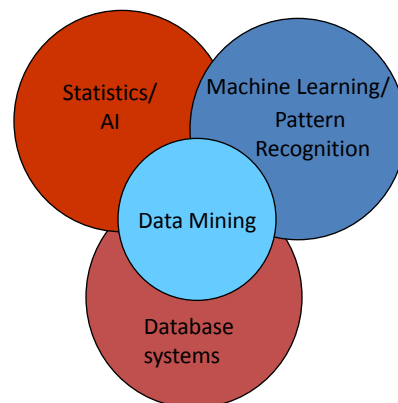


KDD Process: Several Key Steps

- Learning the application domain
 - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**
 - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
 - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
 - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
 - Enormity of data
 - High dimensionality of data
 - Heterogeneous, distributed nature of data



Use (or Misuse) of Statistics on CricInfo (An article published during WC 2011)

- The table below shows that the reviews asked for by batsmen have yielded greater success than those summoned by the fielding team, which is as you'd expect since batsmen usually know when they've been wrongly given out. One out of every three reviews asked by a batsman has been overturned, while that ratio is about one in four for the fielding side.

Overall stats for the DRS				
	Reviews	Successful	Unsuccessful	% successful appeals
Fielding team	55	13	42	23.64
Batting team	41	14	27	34.15
Overall	96	27	69	28.13

Sajjad Haider

Fall 2011

15

Statistics on CricInfo (Cont'd)

DRS for lbws and catches in the World Cup so far				
Dismissal type	Reviews	Successful	Unsuccessful	% successful
Catch	5	1	4	20.00
LBWs - Fielding team	50	12	38	24.00
LBWs - Batting team	41	14	27	34.15
LBWs - Overall	91	26	66	28.57

Sajjad Haider

Fall 2011

16

Best Batsman of Pakistan

- There are always discussions on the best batsman Pakistan has ever produced.
- Most of the times, the opinion is divided between Javed Miandad and Inzimidam-ul-Haq.
- Who is your choice and why?
- Subjectivity vs. Objectivity in your analysis

Best Cricketer of Pakistan (Cont'd)

Test Record	Miandad	Inzamam
Average	52.6	49.6
Home	61.4	53.8
Away	45.8	45.9
1 st Inning	69.1	51.6
2 nd Inning	45.5	60.9
3 rd Inning	39.8	51.7
4 th Inning	54.4	31.7

Best Batsman of Pakistan (Cont'd)

ODI	Miandad	Inzamam
Average	41.7	39.5
Home	48.1	57.7
Away	41.3	38.2
Neutral	41.6	36.8
1 st Inning	42.5	40.0
2 nd Inning	43.9	41.7

Why Not Traditional Data Analysis?

- Tremendous amount of data
 - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
 - Micro-array may have tens of thousands of dimensions
- High complexity of data
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data
 - Structure data, graphs, social networks and multi-linked data
 - Heterogeneous databases and legacy databases
 - Spatial, spatiotemporal, multimedia, text and Web data
 - Software programs, scientific simulations
- New and sophisticated applications

Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
 - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
 - A pattern is [interesting](#) if it is [easily understood](#) by humans, [valid](#) on new or test data with some degree of [certainty](#), [potentially useful](#), [novel](#), or [validates some hypothesis](#) that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
 - [Objective](#): based on [statistics and structures of patterns](#), e.g., support, confidence, etc.
 - [Subjective](#): based on [user’s belief](#) in the data, e.g., unexpectedness, novelty, actionability, etc.

Data Mining Functionalities

- Multidimensional concept description: Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation vs. causality
 - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
 - Construct models (functions) that describe and distinguish classes or concepts for future prediction
 - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
 - Predict some unknown or missing numerical values

Data Mining Functionalities (2)

- Cluster analysis
 - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
 - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
 - Outlier: Data object that does not comply with the general behavior of the data
 - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
 - Trend and deviation: e.g., regression analysis
 - Sequential pattern mining: e.g., digital camera → large SD memory
 - Periodicity analysis
 - Similarity-based analysis
- Other pattern-directed or statistical analyses

Data Mining Tasks

- Prediction Methods
 - Use some variables to predict unknown or future values of other variables.
- Description Methods
 - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

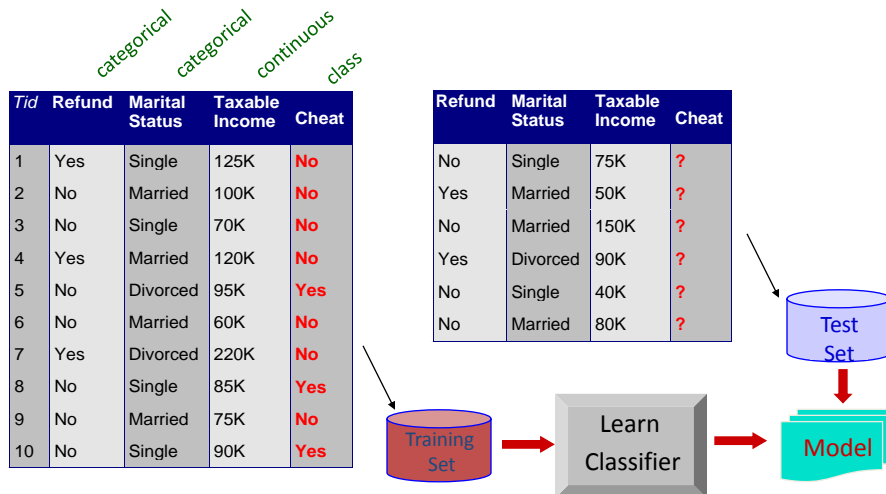
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example



Sajjad Haider

Fall 2011

27

Classification: Application 1

- Direct Marketing
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before.
 - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Sajjad Haider

Fall 2011

28

Classification: Application 2

- Fraud Detection
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Clustering Definition

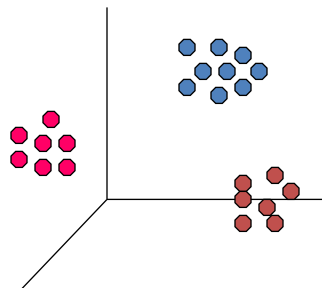
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
 - Data points in one cluster are more similar to one another.
 - Data points in separate clusters are less similar to one another.
- Similarity Measures:
 - Euclidean Distance if attributes are continuous.
 - Other Problem-specific Measures.

Illustrating Clustering

☒ Euclidean Distance Based Clustering in 3-D space.

Intracluster distances
are minimized

Intercluster distances
are maximized



Clustering: Application 1

- Market Segmentation:
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
 - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
 - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}
 {Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
 - Let the rule discovered be
 $\{Bagels, \dots\} \rightarrow \{Potato\ Chips\}$
 - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
 - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
 - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
 - A classic rule --
 - If a customer buys diaper and milk, then he is very likely to buy beer.
 - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

- Inventory Management:
 - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
 - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Regression

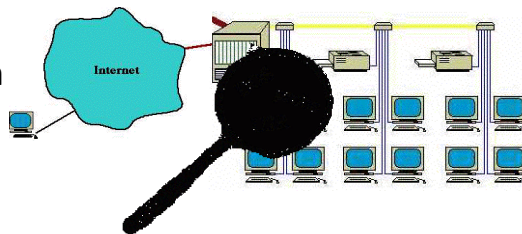
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection



- Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.