Overview of

Data Mining Techniques

Dimitrios Gunopulos, UCR

Nikos Koudas, ATT Research

Outline

- Overview of data mining

- Data Mining Tasks
  - Association Rules
  - Quantitative Association Rules
  - Sequential Patterns
  - Time Series Similarity
  - Clustering
  - Classification

What is Data Mining?

Data Mining:
The <u>efficient</u> discovery of

- previously unknown,
- valid,
- potentially useful,
- understandable

patterns in large databases.

What is Data Mining?

1. Very little functionality in database systems to support mining applications

2. Beyond SQL Querying

   - SQL (OLAP) query:
     - How many widgets did we sell in the 1st Qtr of 1997 in California vs. New York?
   - Data mining query:
     - Which sales region had anomalous sales in the 1st Qtr of 1997?
     - How do the buyers of widgets in California and New York differ?
     - What else do the buyers of widgets in California buy along with widgets?

Overview of the terms

1. Data:

   a set of facts (items) D, typically stored in a database.

2. Pattern:

   an expression E in a language L, that descibes a subset of facts in D.

3. Attribute:

   a field in the description of item $i \in D$.

4. Interestingness:

   a function $I_{D,L}$ that maps an expression $E \in L$ into a measure space M.

Overview of the terms, cont'd

The **Data Mining Task:**

for a given dataset D, language of facts L, interestingness function $I_{D,L}$, and threshold c, find the set of expressions E such that $I_{D,L}(E) < c$ efficiently.

Examples of Large Datasets

1. Government...

2. Large corporations

   - WALMART: 20M transactions per day
   - MOBIL: 100TB geological databases.

3. Scientific

   - NASA, Earth Observing System: 50GB per hour.
   - Environmental datasets.

Examples of Data mining Applications

1. Fraud detection
2. Marketing
3. Data Warehousing
4. Astronomy
5. Molecular biology

How is Data Mining Used

1. Identify the problem.
2. Use data mining techniques to transform the data into information.
3. Act on the information.
4. Measure the results.

The Data Mining Process

1. Understand the domain.

2. Create a dataset:

    - Select the interesting attributes.

    - Data cleaning and preprocessing.

3. Choose the data mining task and the specific algorithm.

4. Interpret the results, and possibly return to 2.

Data Mining Operations

- Associations
- Sequential patterns
- Time-Series Clustering
- Classification
- Segmentation

Associations

- $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$: a set of literals, called items.
- Transaction T : a set of items such that $T \subseteq \mathcal{I}$.
- Database $\mathcal{D}$: a set of transactions.
- A transaction T *contains* X, a set of some items in $\mathcal{I}$, if $X \subseteq T$.
- An *association rule* is an implication of the form $X \Longrightarrow Y$, where $X, Y \subset \mathcal{I}$
- The rule $X \Longrightarrow Y$ holds in the transaction set $\mathcal{D}$ with *confidence* c if c% of transactions in $\mathcal{D}$ that contain X also contain Y.
- The rule $X \Longrightarrow Y$ has *support* s in the transaction set $\mathcal{D}$ if s% of transactions in $\mathcal{D}$ contain $X \cup Y$.

Find all rules that have support and confidence greater than user-specified minimum support and minimum confidence.

Example

Transaction ID	Items Bought
2000	1, 2, 3
1000	1, 3
4000	1, 4
5000	2, 5,6

For Minimum Support = 50%, Minimum Confidence = 50%, we get the rules:

$1 \Longrightarrow 3$ with 50% support and 66.6% confidence.

$3 \Longrightarrow 1$ with 50% support and 100% confidence.

Characteristics

Number of items: Hundreds of Thousands

Number of transactions: Millions

Data set sizes: High Gigabytes

- *Discovering* all rules rather than *verifying* if a rule holds
- Completeness
- Performance
- Scalability

Sample Applications

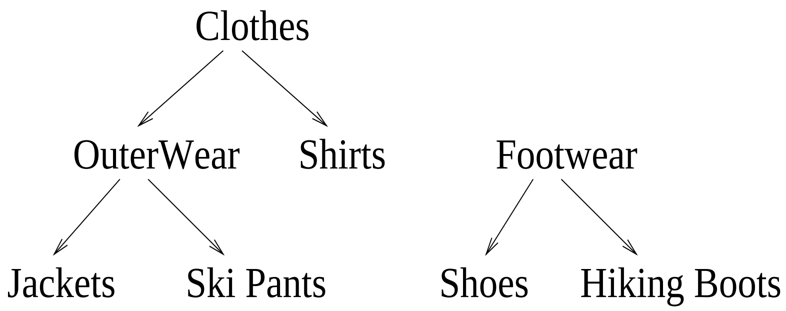
- Market Basket Analysis
  - "$* \Longrightarrow$ Maintenance Agreement"

    What the store should do to boost Maintenance Agreement sales?
  - "Home Electronics $\Longrightarrow *$"

    What other products should the store stock up on if the store has a sale on Home Electronics?
- Attached mailing in direct marketing
- Detecting "ping-pong"ing of patients
  - transaction: patient
  - item: doctor/clinic visited by a patient
  - support of a rule: number of common patients
- HIC Australia "success story"
  - associations between medical payment codes
  - saved $500,000 per year per state

Associations Across Hierarchies

Hierarchies over items (e.g. UPC codes, brand, categories, and product groups)



"Outerwear $\Longrightarrow$ Shoes" may hold, even if

1. "Jackets $\Longrightarrow$ Shoes" does not hold (not enough support)
2. "Clothes $\Longrightarrow$ Shoes" does not hold (not enough confidence)

Rule Understanding

- Rule interest measures to characterize rules into positive, negative, and neutral rules

- Control of false discovery

- Visualization of rules

Quantitative Association Rules

- Given a relational table:
  - a set of records
  - each record has categorical & quantitative attributes
- Example: "30% of married people between age 45 and 60 have at least 2 cars"
- Find all associations supported by more than a user-specified percentage of records.
- Applications: insurance and finance industries

Quantitative Association Rules: Approach

- Fine-partition values of quantitative attribute.
- Combine adjacent partitions as necessary.
- Partial completeness measure quantifies information loss due to partitioning.
- "greater-than-expected-value" interest measure to identify the interesting rules.
  - but also look at "descendants"
- Extend algorithm for boolean associations to mine quantitative rules.

Sequential Patterns

- Given:
  - A database of customer transactions
  - Each transaction is a set of items
- Example: 10% of customers bought "shirts" and "jackets" in one transaction, followed by "shoes" in another transaction.
  - 10% is called the *support* of the pattern
- Find all sequential patterns supported by more than a user-specified percentage of customers.
- Constraints
  - max/min time gap between elements
  - "sliding window" transactions
- Applications:
  - attached mailing
  - customer satisfaction
  - medical research

Sequential Patterns: Example

Database of Customer Transactions Sorted by Customer Id and Transaction Time

Cust. Id	Trans. Time	Items Bought
1	June 2 '93	<u>Shoes</u>
1	Dec 3 '93	<u>Shoes</u>
2	June 1 '93	<u>Shoes</u>
2	June 5 '93	Jacket
2	Sep 2 '93	Shirt, Tie
3	Feb 2 '93	Shoes, Tie
4	June 2 '93	<u>Shoes</u>
4	July 3 '93	Shirt, Belt, Tie
4	Oct 2 '93	<u>Shoes</u>
5	Dec 1 '93	Shoes

Examples of sequential patterns with support ≥ 2 customers:

Shoes, *followed by* Shirt and Tie
<u>Shoes</u>, *followed by* <u>Shoes</u>

Time Series Clustering

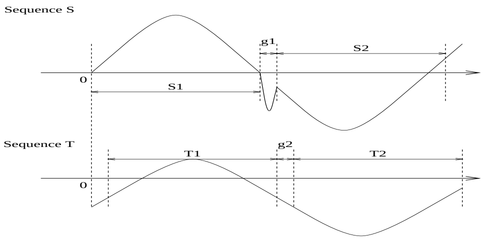
- Given:
  - A database of time-series
- Find:
  - Groups of "similar" time-series

Sample Applications:

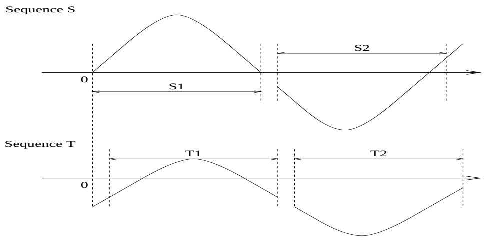
- Find customers with similar pattern of electricity consumption.
- Determine products with similar selling patterns.
- Identify companies with similar pattern of growth.
- Find stocks with similar price movements.

Illustration of Time-series Matching

(1) Original Time-Series

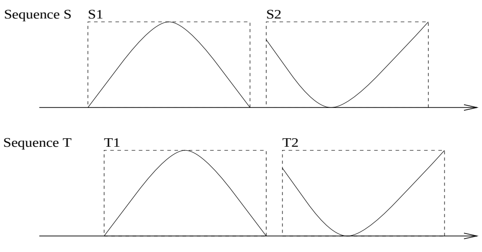


(2) Ignoring Noise



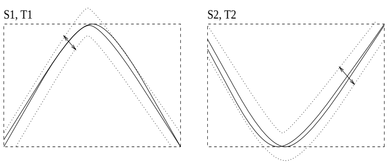
(3) Offset Translation

(3) + Amplitude Scaling

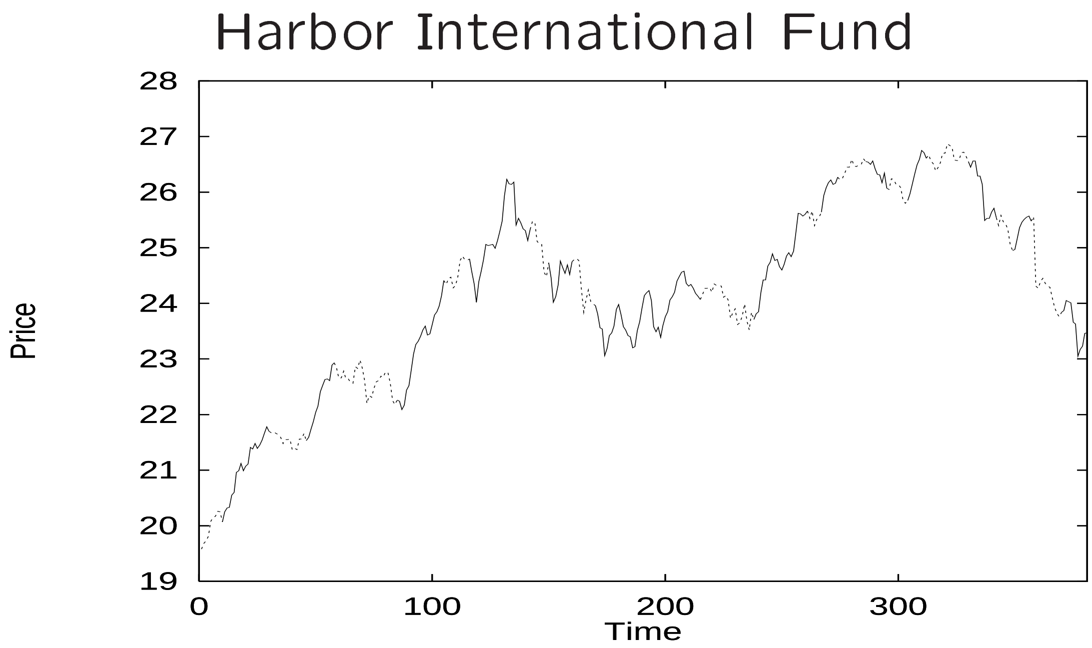


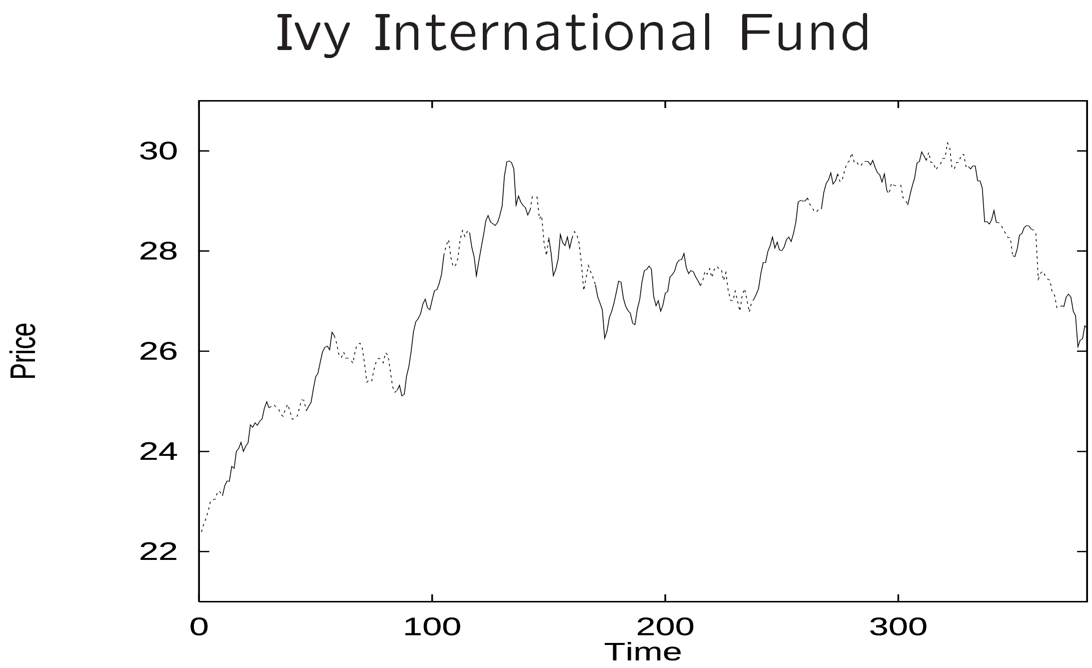
(4) Subseries Matching

(4)



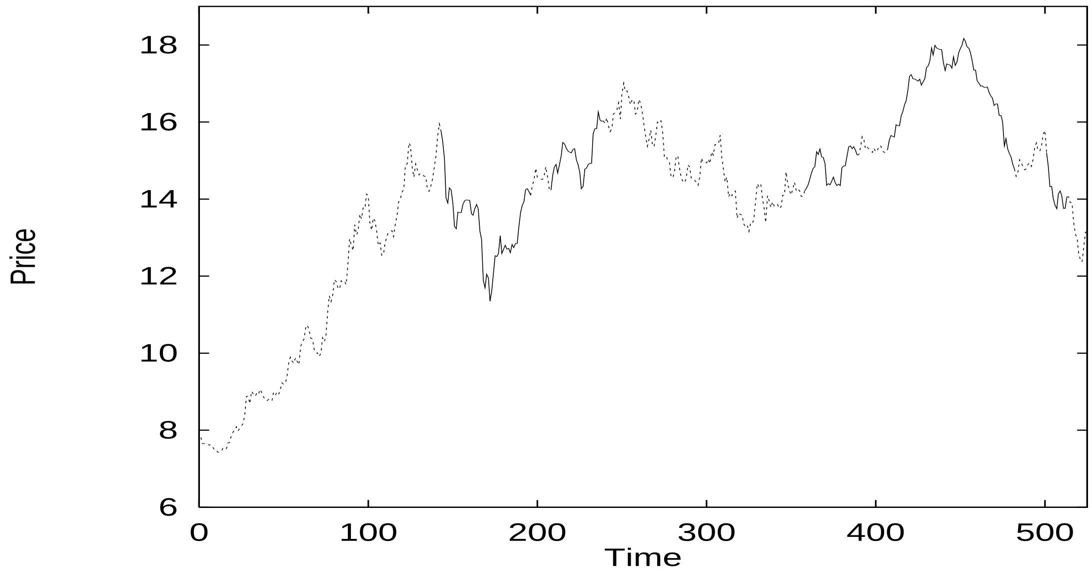
Two Similar Mutual Funds in the Same Fund Group



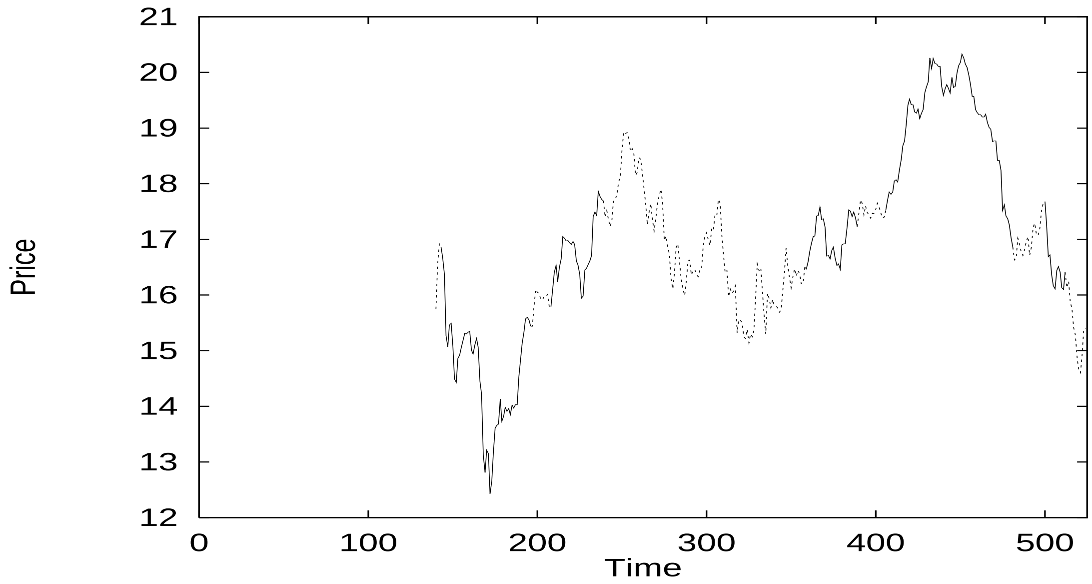


Two Similar Mutual Funds in Different Fund Groups

VanEck International Investor Fund



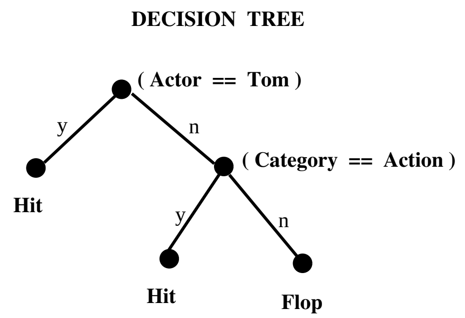
Fidelity Selective Precious Metal and Mineral Fund



Classification

The problem: Given records with class labels, find model for each class

Actor	Category	Class
Tom	Action	Hit
Arnold	Comedy	Flop
Tom	Tragedy	Hit
Bruce	Action	Hit
Bruce	Tragedy	Flop
Arnold	Action	Hit
Tom	Comedy	Hit



Sample Applications:

- Store location
- Treatment effectiveness
- Credit approval
- Target marketing

Issues

- A large variety of solutions exists
  - Tree-based
  - Neural Networks
  - Genetic Algorithms
  - Statistical Algorithms

- Classification Goals
  - Accuracy
    - Several comparable classifiers
    - No clear winner
  - Scalability
    - Many classification algorithms work only with in-memory data
    - Restrictions on number of tuples/attributes
  - Robustness
    - Need to handle both continuous and categorical attributes
  - Interpretability
    - Classifier output should be easy to understand

Clustering

- Given:
  - Data about entities
- Group entities into classes

Sample Applications:

- Customer segmentation
- Clinical trial evaluation

Summary

A new generation of applications:

- Associations
- Sequential patterns
- Time-series clustering
- Classifications
- Segmentation

Combination of intriguing theoretical questions and large commercial interest

Data Mining Research Issues

1. Handling large databases:
   - High dimensionality
   - Scalability
2. Integrating data mining algorithms and DBMSs and SQL.
3. Handling missing data and outliers.
4. Identify and model complex relations between attributes.

Data Mining Research Issues, Cont'd

1. Understandability of results:
   - User interaction.
   - Modeling prior knowledge.
   - Identify the interesting patterns.

2. New applications.
   (a) Finding important web pages.
   (b) Understanding web commmunities.
   (c) Finding motifs in biological sequences.