Introduction to Spatial Data Mining

7.1 Pattern Discovery
7.2 Motivation
7.3 Classification Techniques
7.4 Association Rule Discovery Techniques
7.5 Clustering
7.6 Outlier Detection
Introduction: a classic example for spatial analysis

Disease cluster

Dr. John Snow
Deaths of cholera epidemic
London, September 1854

A good representation is
the key to solving a problem
Good representation because...

- Represents spatial relation of objects
- of the same type
- Represents spatial relation of objects to other objects

Shows only relevant aspects and hides irrelevant.

*It is not only important where a cluster is but also, what else is there (e.g. a water-pump)!*
Other examples of Spatial Patterns

- **Historic Examples** (section 7.1.5, pp. 186)
  - Fluoride and healthy gums near Colorado river
  - Theory of Gondwanaland - continents fit like pieces of a jigsaw puzzle

- **Modern Examples**
  - Cancer clusters to investigate environment health hazards
  - Crime hotspots for planning police patrol routes
  - Bald eagles nest on tall trees near open water
  - Nile virus spreading from north east USA to south and west
  - Unusual warming of Pacific ocean (El Nino) affects weather in USA
Goals of Spatial Data Mining

• Identifying spatial patterns
• Identifying spatial objects that are potential generators of patterns
• Identifying information relevant for explaining the spatial pattern (and hiding irrelevant information)
• Presenting the information in a way that is intuitive and supports further analysis
What is a Spatial Pattern?

• What is not a pattern?
  • Random, haphazard, chance, stray, accidental, unexpected
  • Without definite direction, trend, rule, method, design, aim, purpose
  • Accidental - without design, outside regular course of things
  • Casual - absence of pre-arrangement, relatively unimportant
  • Fortuitous - What occurs without known cause

• What is a Pattern?
  • A frequent arrangement, configuration, composition, regularity
  • A rule, law, method, design, description
  • A major direction, trend, prediction
  • A significant surface irregularity or unevenness
What is Spatial Data Mining?

Metaphors
- Mining nuggets of information embedded in large databases
  - Nuggets = interesting, useful, unexpected spatial patterns
  - Mining = looking for nuggets
- Needle in a haystack

Defining Spatial Data Mining
- Search for spatial patterns
- Non-trivial search - as “automated” as possible—reduce human effort
- Interesting, useful and unexpected spatial pattern
What is Spatial Data Mining? - 2

- Non-trivial search for **interesting** and **unexpected** spatial pattern
- Non-trivial Search
  - Large (e.g. exponential) search space of plausible hypothesis
  - Ex. Asiatic cholera: causes: water, food, air, insects, ...; water delivery mechanisms - numerous pumps, rivers, ponds, wells, pipes, ...
- Interesting
  - Useful in certain application domain
  - Ex. Shutting off identified Water pump => saved human life
- Unexpected
  - Pattern is not common knowledge
  - May provide a new understanding of world
  - Ex. Water pump - Cholera connection lead to the “germ” theory
What is NOT Spatial Data Mining?

- **Simple Querying of Spatial Data**
  - Find neighbors of Canada given names and boundaries of all countries
  - Find shortest path from Boston to Houston in a freeway map
  - Search space is not large (not exponential)

- **Testing a hypothesis via a primary data analysis**
  - Ex. Female chimpanzee territories are smaller than male territories
  - Search space is not large !
  - SDM: secondary data analysis to generate multiple plausible hypotheses

- **Uninteresting or obvious patterns in spatial data**
  - Heavy rainfall in Minneapolis is correlated with heavy rainfall in St. Paul, Given that the two cities are 10 miles apart.
  - Common knowledge: Nearby places have similar rainfall

- **Mining of non-spatial data**
  - Diaper sales and beer sales are correlated in evenings
  - GPS product buyers are of 3 kinds:
    - outdoors enthusiasts, farmers, technology enthusiasts
Why Learn about Spatial Data Mining?

- Two basic reasons for new work
  - Consideration of use in certain application domains
  - Provide fundamental new understanding

- Application domains
  - Scale up secondary spatial (statistical) analysis to very large datasets
    - Describe/explain locations of human settlements in last 5000 years
    - Find cancer clusters to locate hazardous environments
    - Prepare land-use maps from satellite imagery
    - Predict habitat suitable for endangered species
  - Find new spatial patterns
    - Find groups of co-located geographic features
Why Learn about Spatial Data Mining? - 2

- New understanding of geographic processes for Critical questions
  - Ex. How is the health of planet Earth?
  - Ex. Characterize effects of human activity on environment and ecology
  - Ex. Predict effect of El Nino on weather, and economy

- Traditional approach: manually generate and test hypothesis
  - But, spatial data is growing too fast to analyze manually
    - Satellite imagery, GPS tracks, sensors on highways, ...
  - Number of possible geographic hypothesis too large to explore manually
    - Large number of geographic features and locations
    - Number of interacting subsets of features grow exponentially
    - Ex. Find tele connections between weather events across ocean and land areas

- SDM may reduce the set of plausible hypothesis
  - Identify hypothesis supported by the data
  - For further exploration using traditional statistical methods
Interactive Exploratory Analysis

Choropleth maps showing distribution of variable(s) in space

Displays dynamically linked

Parallel Coordinate Plot

Combining spatial and non-spatial displays

Variables selected and manipulated by the user

Powerful for low-dimensional dependencies (3-4)

Scatter Plot
Data Mining: A KDD Process

- **Selection**: Obtain data from various sources.
- **Preprocessing**: Cleanse data.
- **Transformation**: Convert to common format. Transform to new format.
- **Data Mining**: Obtain desired results.
- **Interpretation/Evaluation**: Present results to user in meaningful manner.
Data Mining: Confluence of Multiple Disciplines

- Database Systems
- Statistics
- Machine Learning
- Information Theory
- Algorithms, …, other Disciplines
- Visualization
Primary Data Mining Tasks

- **Descriptive Modeling**
  - Finding a compact description for large dataset
  - Clustering: group objects into groups based on their attributes
  - Association rules: correlate what events are likely to occur together
  - Sequential rules: correlate events ordered in time
  - Trend detection: discovering the most significant changes

- **Predictive Modeling**
  - Classification: assign objects into groups by recognizing patterns
  - Regression: forecasting what may happen in the future by mapping a data item to a predicting real-value variable
What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Intra-cluster distances are minimized

Inter-cluster distances are maximized
Clustering

Cluster: a collection of data objects

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Clustering

- Grouping a set of data objects into clusters based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Example

- Land use: Identification of areas of similar land use in an earth observation database
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
**Association rule**

- **Association (correlation and causality)**
  
  \[ \text{age}(X, \text{“20..29”}) \land \text{income}(X, \text{“20..29K”}) \rightarrow \text{buys}(X, \text{“PC”}) \]  
  [support = 2%, confidence = 60%]

- **Association rule mining**
  
  - Finding frequent patterns, associations, correlations among sets of items or objects in transaction databases, relational databases, and other information repositories
  
  - **Frequent pattern**: pattern (set of items, sequence, etc.) that occurs frequently in a database

- **Motivation**: finding regularities in data
  
  - What products were often purchased together?
Example: Association rule

- Itemset $A_1, A_2 = \{a_1, \ldots, a_k\}$
- Find all the rules $A_1 \Rightarrow A_2$ with min confidence and support
  - support, $s$, probability that a transaction contains $A_1 \cup A_2$
  - confidence, $c$, conditional probability that a transaction having $A_1$ also contains $A_2$.

<table>
<thead>
<tr>
<th>Transaction-id</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$a_1, a_2, a_3$</td>
</tr>
<tr>
<td>20</td>
<td>$a_1, a_3$</td>
</tr>
<tr>
<td>30</td>
<td>$a_1, a_4$</td>
</tr>
<tr>
<td>40</td>
<td>$a_2, a_5, a_6$</td>
</tr>
</tbody>
</table>

Let $\text{min\_support} = 50\%$, $\text{min\_conf} = 50\%$:
- $a_1 \Rightarrow a_3$ (50\%, 66.7\%)
- $a_3 \Rightarrow a_1$ (50\%, 100\%)
Deviation Detection

- **Outlier**: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Periodicity analysis
  - Similarity-based analysis
Classification and Regression

- **Classification:**
  - constructs a model (classifier) based on the *training set* and uses it in classifying new data
  - Example: Climate Classification,…

- **Regression:**
  - models *continuous-valued functions*, i.e., predicts unknown or missing values
  - Example: stock trends prediction,…
Classification (1): Model Construction

**Training Data**

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Assistant Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Professor</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Associate Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Assistant Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Associate Prof</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

**Classification Algorithms**

IF rank = ‘professor’ OR years > 6 THEN tenured = ‘yes’
Classification (2): Prediction Using the Model

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>Assistant Prof</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>Merlisa</td>
<td>Associate Prof</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>George</td>
<td>Professor</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>Joseph</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
</tbody>
</table>

Unseen Data
(Jeff, Professor, 4)

Tenured?

Yes
Classification Techniques

- Decision Tree Induction
- Bayesian Classification
- Neural Networks
- Genetic Algorithms
- Fuzzy Set and Logic
Regression

- Regression is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value

- Methods
  - Linear and multiple regression
  - Non-linear regression

- Regression is different from classification
  - Classification refers to predict categorical class label
  - Regression models continuous-valued functions
Spatial Data Mining

- Spatial Patterns
  - Hotspots, Clustering, trends, ...
  - Spatial outliers
  - Location prediction
  - Associations, co-locations

- Primary Tasks
  - Spatial Data Clustering Analysis
  - Spatial Outlier Analysis
  - Mining Spatial Association Rules
  - Spatial Classification and Prediction

- Example: Unusual warming of Pacific ocean (El Nino) affects weather in USA…
Spatial Data Mining

- Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography, meteorology, etc.

- The main difference: **spatial autocorrelation**
  - the neighbors of a spatial object may have an influence on it and therefore have to be considered as well

- **Spatial attributes**
  - Topological
    - adjacency or inclusion information
  - Geometric
    - position (longitude/latitude), area, perimeter, boundary polygon
Example

- What Kind of Houses Are Highly Valued?—Associative Classification
Example: Location Prediction

• Question addressed
  • Where will a phenomenon occur?
  • Which spatial events are predictable?
  • How can a spatial events be predicted from other spatial events?
    • Equations, rules, other methods,

• Examples:
  • Where will an endangered bird nest?
  • Which areas are prone to fire given maps of vegetation, draught, etc.?
  • What should be recommended to a traveler in a given location?
**Example: Spatial Interactions**

- **Question addressed**
  - Which spatial events are related to each other?
  - Which spatial phenomena depend on other phenomenon?

- **Examples:**

<table>
<thead>
<tr>
<th>Domains</th>
<th>Example Features</th>
<th>Example Co-location Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecology</td>
<td>Species</td>
<td>(Nile crocodile, Egyptian plover)</td>
</tr>
<tr>
<td>Earth science</td>
<td>climate and disturbance events</td>
<td>(wildfire, hot, dry, lightning)</td>
</tr>
<tr>
<td>Economics</td>
<td>industry types</td>
<td>(suppliers, producers, consultants)</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>disease types and environmental events</td>
<td>(West Nile disease, stagnant water sources, dead birds, mosquitoes)</td>
</tr>
<tr>
<td>Location-based service</td>
<td>service type requests</td>
<td>(tow, police, ambulance)</td>
</tr>
<tr>
<td>Weather</td>
<td>fronts, precipitation</td>
<td>(cold front, warm front, snow fall)</td>
</tr>
<tr>
<td>Transportation</td>
<td>delivery service tracks</td>
<td>(US Postal Service, UPS, newspaper delivery)</td>
</tr>
</tbody>
</table>

- **Exercise:** List two interaction patterns.
**Example: Hot spots**

- **Question addressed**
  - Is a phenomenon spatially clustered?
  - Which spatial entities or clusters are unusual?
  - Which spatial entities share common characteristics?

- **Examples:**
  - Cancer clusters [CDC] to launch investigations
  - Crime hot spots to plan police patrols

- **Defining unusual**
  - Comparison group:
    - neighborhood
    - entire population
  - Significance: probability of being unusual is high