Spatial Data Mining: Progress and Challenges

Krzysztof Koperski, Jiawei Han, Junas Adhikary
School of Computing Science
Simon Fraser University

Presented by Mufida Shirif
Nov 22, 2000

Contents

- Introduction
- Methods for Knowledge Discovery in Spatial Databases
- Future Directions
- Conclusion

Introduction

- Spatial Data Mining Background
  - Primitives of Spatial Data Mining
  - Spatial Data Structures, Computations and Queries
- Spatial Data Mining Architecture

What is ...

spatial data?
spatial database?
Spatial data mining?

- Spatial data are the data related to objects that occupy space.
  - Geometric property (I.e. spatial location, area,... etc)
  - Topological property (I.e. adjacency, inclusion,... etc)
- A spatial database stores spatial objects represented by spatial data types and spatial relationships among such objects.
- Spatial data mining refers to the extraction of implicit knowledge, spatial relations, or patterns not explicitly stored in spatial databases.

Spatial Data Mining Background

- **Statistical spatial analysis**
  - Handles very well numerical data and usually comes up with realistic models of spatial phenomena
- ***Disadvantages***
  - The assumption of statistical independence among the spatially distributed data
  - Can not model nonlinear rules very well and symbolic values like names are handled poorly
  - Statistical methods do not work well with incomplete data
  - Expensive computation of the result

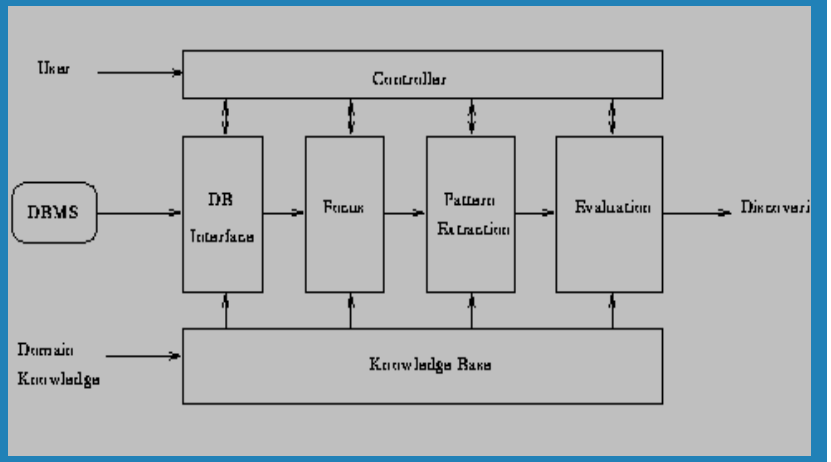
Primitives of Spatial Data Mining

- **Rules:**
  - **A spatial characteristic rule**
  - **A spatial discriminant rule**
  - **A spatial association rule**
- **Thematic Maps**
  - **Raster (represented in the form as the intensity of the pixels)**
  - **Vector (represented by its geometry)**
- **Image Databases**
  - **Data consists of images or pictures**

Spatial Data Structure, Computations, and Queries

- **Spatial Data Structures**
  - **Using R - tree**
- **Spatial Computations**
  - **Spatial join**
- **Spatial Query Processing**
  - **SAND (An architecture for spatial databases)**

Spatial Data Mining Architecture



Methods for Knowledge Discovery in Spatial Data

- Generalization-Based
- Methods using Clustering
- Methods exploring spatial association
- Using Approximation and Aggregation
- Mining in Image and Raster Databases

Generalization-Based Knowledge Discovery

- Requires background knowledge in the form of concept hierarchies
  - Spatial
  - Non-spatial
- Two generalization based algorithms
  - Spatial-data-dominant Generalization
  - non-Spatial-data-dominant Generalization
- Disadvantages:
  - Dependent upon the given Concept hierarchies

Methods Using Clustering

- SD (CLARANS)
- NSD (CLARANS)
- Both algorithms assume that the user specifies the type of the rule to be mined and relevant data through a learning request
- Drawback:
  - Objects to be clustered are all stored in memory

Methods Using Clustering (Con't)

- CLARANS in large databases: Focusing Method
  - Focusing on representative objects to reduce the number of objects to consider
  - Restrict the access to certain objects that do not actually contribute to the computation
    - Focusing on relevant clusters
    - Focusing on a cluster
- Results shows that the effectiveness decreased and the efficiency increased when the focusing technique was used

Methods Using Clustering (Con't)

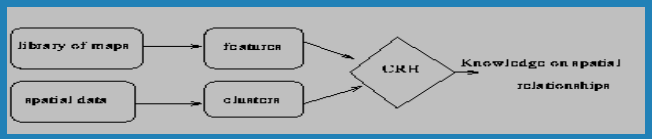
- DBSCAN
  - Can deal with CLARANS problems
- Birch
  - Adjustment of memory requirement to the size of memory that is available
  - The author used concepts called:
    - Clustering Feature and CF trees

Methods Exploring Spatial Associations

- A spatial association rule is of the form X --> Y(c%), where X and Y are sets of spatial or non-spatial predicates and c% is the confidence of the rule
  - Example: is_a(x,school) --> close_to(x,park) (80%)
    - states that 80% of schools are close to parks
- A strong rule is a rule with large support and large confidence
- Algorithm for Multiple Level Spatial Association Rule
  - Collect s the task-relevant data
  - Extract spatial associations at the level of generalized spatial relation using efficient computations
- This algorithm requires background knowledge in the form of concept hierarchies
- Expect a user to describe the form of the rule s/he wants by giving such information in the mining query

Using Approximation and Aggregation

- Aggregate Proximity
  - The aggregate proximity is the measure of closeness of the set of points in the cluster to a feature as opposed to the distance between a cluster boundary and the boundary of a feature
- Algorithm CRH
  - Uses as filters to reduce the candidate features at multiple level



  - The problem is that a feature may have to be tested for overlap with a cluster many times.
- Finding Commonalties and Maximal Discriminators between Clusters

Mining in Image and Raster databases

- Data Mining in image databases may be seen as similar to image processing
- Decision tree method was used
- Example:
  - Identifying volcanoes on the surface of Venus from images transmitted by the Magellan spacecraft
- CONQUEST was use for raster databases mining
  - Detecting earthquakes from space

Future Directions

- Data Mining in spatial object-oriented databases
- Alternative clustering techniques
- Mining spatial data deviation and revolution rules
- Using multiple thematic maps
- Parallel data mining
- Mining of remotely sensed images
- Multidimensional rule visualization
- Intelligent GIS methods

Conclusion

- Spatial data mining is a promising field of research with a wide applications in GIS, medical imaging, robot motion planing, ...etc.
- A number of algorithms and techniques have been proposed to discover various kind of knowledge from spatial data