Spatial Data Mining: Progress and Challenges Survey Paper

Krzysztof Koperski, Junas Adhikary, and Jiawei Han (1996)

Review by Brad Danielson
CMPUT 695
01/11/2007

Introduction

Authors objectives:

- Describe and critique existing spatial data mining methods
- Give readers a general perspective of the field's current state
- Make suggestions for future directions and growth potential of spatial data mining

Introduction

My objectives:

- Summarize the paper's description of the state of spatial data mining in 1996.
- Examine the predictions for future directions made by these authors.
- Briefly examine the accuracy of these predictions by doing a topic search on spatial data mining research from 1997 to 2007.

Spatial Data Mining: Definition

- "*Spatial data mining, or knowledge discovery in spatial database, refers to the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases." (Koperski and Han, 1995)*
- Data mining, or knowledge discovery in databases, refers to the *" discovery of interesting, implicit, and previously unknown knowledge from large databases." (Frawley et al, 1992)*

WHAT'S THE DIFFERENCE?

What is spatial data and why is mining it different that mining "normal" data?

DATA = *The (WHAT) dimension.*

an attribute of an object.

SPATIAL DATA = (WHERE) & (WHAT)

Attribute data referenced to a specific location.

The Attributes of spatial objects are:

- Highly dependant on location
- Often influenced by neighboring objects

Spatial Data …?

Object	Where	What
Milk	In fridge (X_1,Y_1)	Is cold
Milk	On table (X_2,Y_2)	Is warm

SPATIAL DATA MINING:

THE FRIDGE IS BROKEN!

Object	Where	What
Milk	In fridge (X_1,Y_1)	Is warm
Yogurt	In fridge (X_1,Y_1)	Is warm
Butter	In fridge (X_1,Y_1)	Is warm

Object() : Location() -> Characteristic()

(Milk) : (In Fridge) -> [should be] -> (Cold)

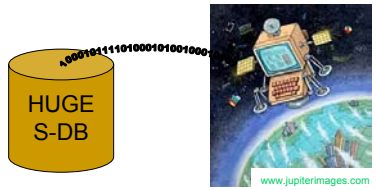
Why do Spatial Data Mining?

- To understand spatial data
- To discover relationships between spatial and non spatial data
- To capture the general characteristics in a concise way
- To build spatial knowledge-bases

Critical Challenge in Spatial Data Mining

SD mining algorithms must efficiently overcome:

- The huge volume of spatial data
- The complexity of spatial data types/structures
- The complexity of spatial accessing/query methods
- Expensive spatial processing operations



Object =

Which highways cross Nat'n Park boundaries?

= spatial JOIN

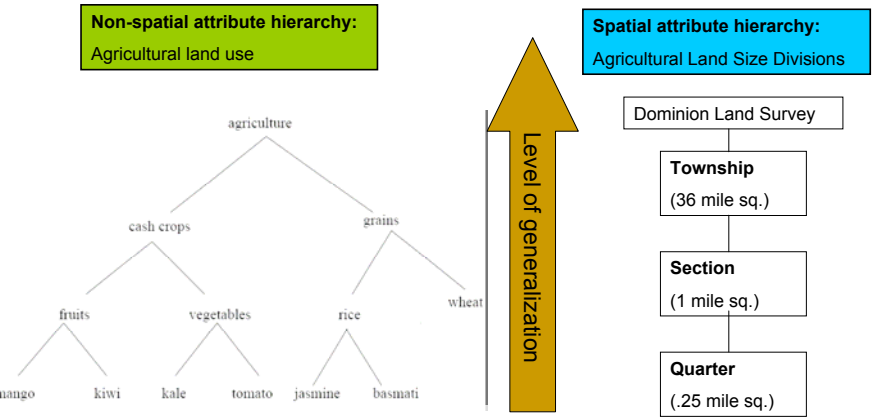
Spatial Data Mining Methods

1. Generalization Based Knowledge Discovery
2. Clustering Methods
3. Aggregate Proximity Measuring
4. Spatial Association Rules

Generalization-Based Knowledge Discovery

- Requires background knowledge of dataset, presented as *concept hierarchies*
- Developing these hierarchies conceptually similar to Hierarchical Clustering
- Hierarchies are either based on **spatial attributes** or **non-spatial attributes**.
- Queries about data can then be made at levels of generalization in the hierarchy.

Concept Hierarchies



Queries can be made at different levels of Generalization

Spatial vs Non-spatial Dominant Generalization

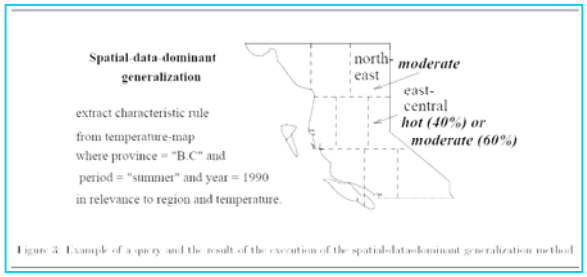
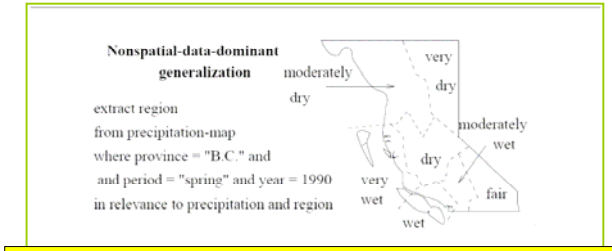


Figure 3. Example of a query and the result of the execution of the spatial-data-dominant generalization method

Spatial objects ***(individual regions)*** are generalized *(merged)* until the desired zone size is reached.

Non-spatial data falling into these zones is then generalized and reported or



Discrete precipitation values generalized into 7 classes.

Spatial data were classified based on their fit to these precipitation attribute classes.

Goal: produce high-level descriptions of the data. Thematic maps are effective

Clustering Methods

- Goal: like Generalization, to reveal relationships between spatial and non-spatial attributes
- Techniques used are based on some clustering methods we examined in class:
  - PAM (k–medoids clustering)
  - CLARA (k-medoids, where medoids are chosen from a sample of a large DB)
  - CLARANS (mixture of PAM and CLARA, where new k-medoids are tested from new random samples)
- 2 spatial data mining variations of CLARANS:
  - SD(CLARANS): spatial dominant approach
  - NSD(CLARANS): non-spatial dominant approach

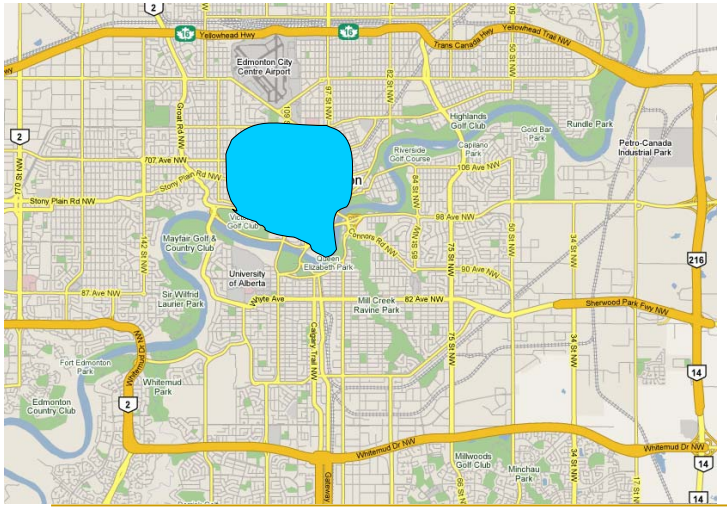
SD(CLARANS)

1. The spatial components of objects in the dataset are collected and clustered using CLARANS
2. Non-spatial description of the objects are brought into the resulting clusters.
3. Result: each cluster (defined by spatial boundaries) is described by it's relative abundance of non-spatial attributes.

NSD(CLARANS)

1. Non-spatial attribute generalization produces *k* generalized attribute groups
2. The spatial components are clustered using CLARANS to find k clusters.
3. If these spatial clusters overlap, they may be merged, and their attribute descriptions merged as well.
4. Result: each cluster is described by a single attribute description

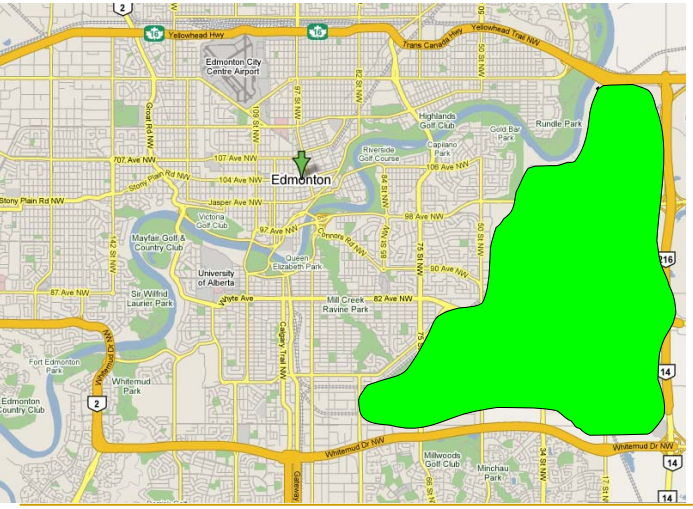
SD(CLARANS): Example



Spatial Cluster: Downtown Edmonton

Attributes: (Building types) 50% Commercial, 40% Residential, 10% Public Service

NSD(CLARANS): Example



Attribute Cluster (Building types): Mostly Industrial

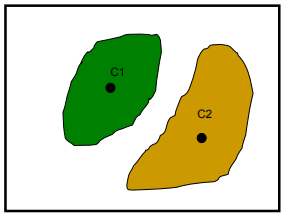
Spatial Cluster: Region East of 50St & South of HW#16 & North of Whitemud &…

Clustering Methods: Improvements

- CLARANS is inefficient at calculating total distance between clusterings.
  - Integrate with more efficient spatial access methods developed for Spatial Databases
- Use cluster focusing to increase efficiency in finding locally optimized clusters:
  - Information about sub-clusters is summarized in tree structures
  - Sub-clusters (or nodes in the sub-cluster tree) can be tested when selecting a new medoid (center) during the recursive process of optimizing clusters
  - (mixture of Hierarchical Clustering and Partition Clustering)

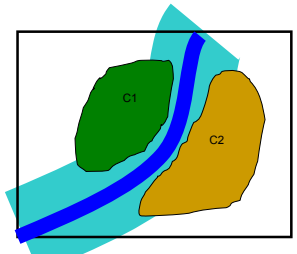
Aggregate Proximity Measuring

- Clustering methods are effective at finding *where* groups of data are in a spatial DB
- BUT – it's often more interesting to know *WHY* the clusters are there.



Cluster C1 centered at X_1, Y_1

Cluster C2 centered at X_2, Y_2

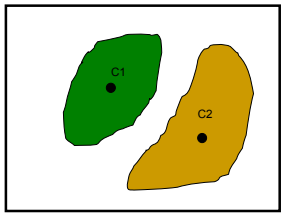


45% of the objects in Clusters C1 & C2 are close to feature{River}

Aggregate Proximity Measuring

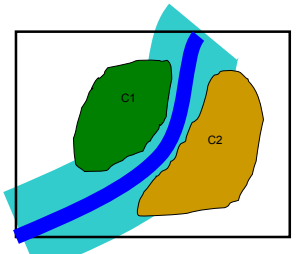
Calculates the distance between a feature and the set of points in a cluster.

This may reveal how outside features influence the cluster.



Cluster C1 centered at X_1, Y_1

Cluster C2 centered at X_2, Y_2

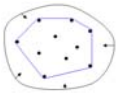


45% of the objects in Clusters C1 & C2 are close to feature{River}

Aggregate Proximity Measuring – How?

CRH Algorithm (**C**ircle, **R**ectangle, convex **H**ull)

- **Input:** any number of local map features AND one cluster
- Uses filters of various geometries to find the characteristics of a cluster wrt nearby features
- The **C** and **R** filters prune candidate features, and only promising features are sent to the **H** filter.
- The **H** filter builds more accurate buffers around the selected features.



- Aggregate proximity is calculated between the points in the cluster and the buffers around the features of interest.
- **Output:** list of features with smallest aggregate proximity to points in the cluster, and percentages of points located within a defined distance from features.

Mining Spatial Association Rules

- Generalization and Clustering characterize spatial objects based on their non-spatial attributes.
- Spatial Association Rules are required to associate spatial objects with other spatial objects.

- Examples from paper:

is_a(x,school) -> close_to(x, park)(80%)

Reads:

80% of schools are close to parks.

- Suggested predicates for spatial association rules:
  - topological relations: *intersects, overlap, disjoint*
  - spatial orientations: *left_of, west_of, etc.*
  - distance information: *close_to, far_from, etc*

min support and *min confidence* are required to filter out infrequent and weak rules.

Multi-level Mining of Spatial Association Rules

- Query: describe a set of objects using relations to other objects. *is_a(x,school) -> adjacent_to(x, park)*
- High Level mining:
  - Use coarse spatial predicates such as *g_close_to (generalized close to)*
  - Spatial objects satisfy *g_close_to* if the distance between their *Minimum Bounding Rectangles* is less than a threshold.
- Low Level mining:
  - More detailed, accurate predicates are used to build rules with objects which pass through from High Level.
- Apriori rational: if a pattern is not large at High Level, it will not be large at Low Level
  - This minimizes expensive spatial computations

Predictions, circa 1996

"The variety of yet unexplored topics and problems makes knowledge discovery in spatial databases an attractive and challenging research field."

- Identified Future Directions for spatial data mining:
  - Data Mining in Spatial Object Oriented DB
  - Alternative Clustering Techniques:
    - Clustering overlapping objects, Fuzzy Clustering of Spatial Data
  - Mining under uncertainty:
    - Evidential reasoning, Fuzzy sets approaches
  - Spatial Data Deviation and Evolution Rules:
    - Rule application to data that changes over time
  - Interleaved Generalization (spatial and non-spatial)
  - Generalization of Temporal Spatial Data (data evolution)
  - Parallel Data Mining (multi-processor systems)
  - Spatial Data Mining Query Language
  - Multidimensional Rule Visualization and Multiple Thematic Maps

Prediction Accuracy, 10 years later

Topic	Currently Active Research Field	References
DM in Spatial Obj-Oriented DB	Yes	>10 (1997 – 2007)
Alternative / Fuzzy Spatial Clustering	Yes	*1996*, >10 (1997 – 2007)
Mining under uncertainty	Yes – esp. Robo nav	>10 (1997 – 2007)
Deviation / Evolution Rules	Yes – mixed topics	>10 (1997 – 2007)
Interleaved Generalization	Vague…	
Generalization of Temporal Spatial Data	Yes	>10 (1997 – 2007)
Parallel Data Mining	merged	
Spatial Data Mining Query Language	Yes: GeoMiner, GMQL, Spatial SQL	1991, 1994, 1996, 1997 (h&k), >10 (1997 – 2007)
Visualization Topics	Yes – esp. GIS	>10 (1997 – 2007)

Conclusion

- Data Mining / Knowledge Discovery of Spatial Data is a large, active research area.
- While it was a "young" field at the time this survey paper was written, it is quickly maturing in applications such as:
  - Geographic Information Systems
  - Medical Imaging
  - Robotics Navigation

References

- Krzysztof Koperski, Junas Adhikary, Jiawei Han. Spatial Data Mining: Progress and Challenges Survey Paper. Workshop on Research Issues on Data Mining and Knowledge Discovery, 1996
- K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int'l Symp. on Large Spatial Databases (SSD'95)*, pages 47--66, Portland, Maine, Aug. 1995.
- W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 1--27.
- Roddick, Hornsby, and Spiliopoulou. *An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining Research*, TSDM2000, LNAI 2007, pp. 147–163, 2001.c Springer-Verlag Berlin Heidelberg 2001