IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 10, October 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569| A Monthly Peer Reviewed & Referred Journal |

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1110116 |

Integration of Data Mining Techniques with Geographic Information Systems (GIS) for Enhanced Spatial Analysis of Geographic Data

Savitri Kulkarni[1], Ramesh B[2], Nagashree R A[3], Gangappa Demannavar[4], Vinutha H M[5]

Assistant Professors, Department of Computer Science and Engineering, City Engineering College, Bengaluru, Karnataka, India[1,2,3,4,5]

ABSTRACT: This paper explores the integration of data mining techniques with Geographic Information Systems (GIS) to enhance spatial analysis of geographic data. The research begins by delving into the data mining functions specifically designed for geographic data and contrasts them with traditional data mining methods. Geographic data, characterized by spatial relationships and geographical coordinates, demands specialized analytical approaches that differ from conventional data mining techniques. The paper then reviews two predominant research approaches in this domain: the first approach focuses on learning within spatial databases, leveraging large-scale geographic datasets to uncover patterns and trends. The second approach is rooted in spatial statistics, emphasizing the application of statistical methods to spatial data for analytical insights. Through an in-depth comparison of these approaches, we highlight the unique benefits and limitations of each method, while also exploring the commonalities that unify them in their goal to extract meaningful insights from geographic data. These findings, emphasizing the key distinctions between the two approaches, such as the role of spatial relationships in determining patterns. It also identifies shared elements, such as the use of geographic data structures and algorithms, offering a comprehensive perspective on how data mining enhances GIS-based spatial analysis.

KEYWORDS: Spatial Data Mining, Spatial Databases, Rules Induction, Spatial Statistics, Spatial Neighborhood

I. INTRODUCTION

The rise in map production is generating vast amounts of data that are challenging to analyze. To address this, applying data mining techniques to spatial data is increasingly relevant. Unlike traditional data mining, spatial data mining must consider spatial relationships between objects. Spatial data mining is used in decision-making areas such as Geo marketing, environmental studies, and risk analysis. For example, Geo marketing can help retailers determine trade areas and analyze customer profiles based on geographic data. In our project, spatial data mining is applied to traffic risk analysis, using historical accident data combined with information on the road network, population, and buildings. The aim is to identify high-risk areas and understand these risks in their geographic context. Geographic data analysis traditionally relies on conventional statistics and multidimensional methods, which often overlook spatial aspects. Spatial statistics, by contrast, considers the interdependence of nearby observations, forming a distinct field of study. This includes techniques like Geo statistics, Exploratory Spatial Data Analysis (ESDA), and the Geographical Analysis Machine (GAM). Recent advancements have extended multidimensional methods to include spatial contiguity, integrating spatial statistics into data mining. Two notable research teams have made significant contributions: the DB Research Lab at Simon Fraser University with Geo Miner, and Munich University with various algorithms and clustering methods. Additionally, the University of Laval has developed data warehouses for spatial data. This paper will outline data mining methods for Geographic Information Systems, focusing on their application in spatial data analysis. It will cover statistical and database-based approaches, compare their similarities and differences, and discuss current research issues.

II. DEFINITION OF SPATIAL DATA MINING

Spatial Data Mining (SDM) involves extracting knowledge, identifying spatial relationships, and uncovering properties not explicitly stored in databases. It aims to discover implicit patterns and relationships between spatial and non-spatial data. The distinct feature of SDM is its focus on spatial interactions. A geographical database represents a spatio-temporal continuum where the attributes of a location are often interconnected with its surroundings. This emphasizes the importance of spatial relationships in analysis. While temporal aspects of spatial data are significant, they are often

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569| A Monthly Peer Reviewed & Referred Journal |

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1110116 |

overlooked. Traditional data mining methods are not well-suited for spatial data because they do not accommodate location data or the implicit relationships between objects. Thus, new methods are needed to handle spatial relationships and data. Calculating these relationships is resource-intensive, generating large volumes of data and potentially impacting performance. Geographic Information Systems (GIS) allow users to query spatial data and conduct basic analytical tasks but are not designed for complex data analysis or knowledge discovery. GIS tools are effective for data access, spatial joins, and graphical map displays, but they lack generic methods for deep analysis and rule inference. Integrating traditional data mining with spatial data mining methods is essential. While conventional data mining focuses on alphanumeric properties, incorporating spatial data mining methods can enhance the analysis and understanding of spatial relationships.

III. SPATIAL DATA MINING TASKS

As detailed in the table below, spatial data mining (SDM) tasks are an extension of conventional data mining tasks, incorporating spatial data and criteria. These tasks are designed to: (i) summarize data, (ii) find classification rules, (iii) create clusters of similar objects, (iv) identify associations and dependencies to characterize data, and (v) detect deviations by analyzing general trends. Different methods are used for these tasks, with some derived from statistics and others from machine learning.

SDM Tasks	Statistics	Machine Learning
Summarization	Global autocorrelation	Density analysis
Characteristic Rules	Class identification	Spatial classification
Clustering	Point pattern analysis	Geometric clustering
Dependencies	Local autocorrelation	Correspondence analysis
Trends and Deviations	Kriging	Trend rules

The following sections will describe the data mining tasks specific to Geographic Information Systems (GIS).

3.1. Spatial Data Summarization

The main goal is to describe data in a global way, which can be done in several ways. One involves extending statistical methods such as variance or factorial analysis to spatial structures.

3.1.1. Statistical analysis of contiguous objects

3.1.1.1. Global Autocorrelation

Summarizing datasets traditionally involves basic statistics such as averages and variances, along with graphical tools like histograms and pie charts. In the context of spatial data, new methods have been developed to measure global neighborhood dependencies, including local variance, local covariance, Geary's spatial autocorrelation, and Moran's indices. These methods utilize a contiguity matrix to represent spatial relationships between objects, which can encompass various types of spatial interactions such as adjacency or distance-based connections.

3.1.1.2. Density Analysis

Density analysis, a component of Exploratory Spatial Data Analysis (ESDA), does not rely on prior knowledge about the data. This technique estimates density by calculating the intensity within small circular windows across a space and then visualizes the point patterns, effectively serving as a graphical method for exploring spatial data.

3.1.1.3. Smooth, Contrast, and Factorial Analysis

Density analysis often ignores non-spatial properties, but geographic data analysis must consider both alphanumerical attributes and spatial data. This requires integrating spatial information with attributes and employing multidimensional analysis methods to examine multiple attributes together. Two primary techniques use the contiguity matrix to integrate spatial neighborhoods into attribute analysis: smoothing, which replaces each attribute value with the average of its neighbors to highlight general data characteristics, and contrasting, which emphasizes variations by subtracting the average from each value. For analyzing multiple attributes, factorial analysis methods are essential. These methods

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569| A Monthly Peer Reviewed & Referred Journal |

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1110116 |

reduce the number of variables by identifying factorial axes with the greatest dispersion of data values. By projecting and visualizing the dataset along these axes, correlations or dependencies between attributes can be identified. Traditional statistical methods assume object independence, but research into spatial organization has led to adaptations of factorial analysis methods to account for spatial contiguity. This involves applying techniques such as Principal Component Analysis or Correspondence Analysis after transforming the data with smoothing or contrasting methods.

3.1.2. Generalization

Generalization involves raising the abstract level of non-spatial attributes and reducing geometric detail by merging adjacent objects. This method extends the concept of attribute-oriented induction, where a concept hierarchy may be spatial (like administrative boundaries) or thematic (like agricultural classifications). For instance, a thematic hierarchy in agriculture might classify "cultivation type" into categories such as "food" and further into "cereals" like maize and wheat. Hierarchies can be either predefined by experts or generated through inference processes, with spatial hierarchies potentially emerging from geometric splits like quad-trees or spatial clustering. There are two types of generalization: non-spatial dominant generalization, which uses thematic hierarchies first before merging adjacent objects, and spatial dominant generalization, which starts with spatial hierarchies and then generalizes non-spatial values for each spatial category. The complexity of algorithms for these processes is O(NlogN), where N represents the number of objects. This approach can serve as a preliminary step for inferring rules, such as association or comparison rules.

3.1.3. Characteristic Rules

Characteristic rules describe properties that are typical for a specific subset of the database but not for the entire database. In spatial databases, these rules consider not only the properties of the objects but also those of their neighboring objects up to a specified level. For a subset S of objects, parameters include significance (relative frequency in S compared to the database), confidence (ratio of objects in S meeting the significance threshold in their neighborhood), and maximum extension (max-neighbors). The result is expressed as rules, such as:

$$S \Rightarrow p\ 1\ (n\ 1\ ,\ freq\text{-}fac\ 1\ ) \wedge \ldots \wedge p\ k\ (n\ k\ ,\ freq\text{-}\ fac\ k\ ).$$

where each rule indicates properties and their frequency factors extended to neighbors.

3.2. Class Identification

Class identification, or supervised classification, provides a logical description for the optimal partitioning of a database. Classification rules, typically represented as decision trees, use criteria that may include spatial predicates, extending beyond just the properties of individual objects to include their neighbors' properties. In spatial statistics, classification often involves analyzing remotely-sensed data to categorize pixels and aggregate homogeneous ones into geographic entities. In spatial databases, classification considers both an object's properties and its neighbors' properties, potentially extending to multiple degrees of neighbors. For example, a classification rule might be:

High population ^ neighbor = road ^ neighbor of neighbor = airport => high economic power (95%).

Geo Miner enhances this approach by integrating spatial attributes, enabling classification based on spatial relationships such as proximity to densely populated areas.

3.3. Clustering

Clustering, an unsupervised classification task, partitions a dataset based on similarity functions. In spatial databases, clustering methods often employ similarity functions such as Euclidean distance to group neighboring objects. Although clustering methods for spatial databases are not significantly different from those for relational databases, they focus on optimizing algorithms for geometric clustering. Geo Miner combines geometric clustering with generalization based on non-spatial attributes, creating new classes that represent areas, like the grouping of major U.S. cities.

Algorithms like CLARANS, DBSCAN, and STING are commonly used for clustering. GDBSCAN, a more recent method, extends these algorithms to spatial data, handling various spatial shapes and incorporating attributes.

3.4. Spatial Data Dependencies

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569| A Monthly Peer Reviewed & Referred Journal |

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1110116 |

Spatial data dependencies can be assessed using local autocorrelation and association rules adapted for spatial data. Local autocorrelation evaluates spatial dependence by comparing actual spatial distributions with random ones, using a spatial weight matrix. Association rules, originally used for market analysis, are extended to spatial data to identify frequent spatial relationships between objects. For example:

$$is\_a\ (x, gas\_station) \wedge within\ (x, rural\_area) \rightarrow close\_to\ (x, highway)\ [65\%, 80\%]$$

This rule indicates that gas stations in rural areas are typically near highways. The Geo Miner algorithm enhances this by evaluating spatial predicates in two phases: an approximate test followed by an exact test.

3.5. Trend and Deviation Analysis

Trend and deviation analysis in spatial databases aims to identify and characterize spatial trends. Using methods based on central place theory, this analysis involves discovering centers, determining theoretical trends, and analyzing deviations. Examples include analyzing unemployment rates relative to proximity to a metropolis or tracking housing development. Geo statistics, originally for geological applications, is now used for spatial and spatio-temporal trend analysis. It employs techniques like kriging to predict values outside sample points. Geo statistics is particularly effective for point set analysis and dealing with single variables or attributes.

IV. COMPARISON OF SDM APPROACHES

This study compares spatial data mining (SDM) methods developed by the Statistics and Database communities, highlighting their similarities and differences. Methods vary between graphical and semantic approaches. Graphical methods, like density and clustering analysis, provide visual insights, whereas semantic methods use graphs and matrices for detailed spatial relationship analysis. The use of contiguity differs between approaches: the learning approach represents spatial relationships as distinct properties, while the statistical approach integrates these relationships into formulas or adjusts initial data. Statistical methods focus on intra-thematic relationships, whereas learning methods also address inter-thematic relationships, crucial for explanatory models. Methods like generalization and factorial analysis offer complementary approaches. Generalization simplifies data for further analysis, while factorial analysis reveals underlying correlations. Combining these methods enhances data preparation and interpretation, such as generalizing before characterization or classification.

This study aims to compare spatial data mining (SDM) methods developed by both the Statistics and Database communities, offering a detailed exploration of their similarities and differences. Spatial data mining refers to extracting patterns and useful information from spatial datasets, and different methodologies exist within these two fields to address this challenge. While the goals of both approaches are similar—understanding spatial relationships and structures within data—the methods used by each community reflect different conceptual frameworks.

The first key distinction lies in the types of methods used, which can be broadly categorized into graphical and semantic approaches. Graphical methods, such as density and clustering analysis, focus on providing visual representations of spatial data. These methods allow researchers to easily observe and interpret patterns, making them particularly useful for initial data exploration and hypothesis generation. On the other hand, semantic methods take a more analytical approach, employing graphs and matrices to offer a more detailed and precise analysis of spatial relationships. These methods are typically more abstract and rely on mathematical representations, which help to uncover more nuanced connections in the data that may not be immediately apparent through visualization alone.

Another critical difference between the two approaches is the treatment of contiguity, or the way spatial relationships are represented. In learning-based approaches, which often come from the database community, spatial relationships are explicitly encoded as distinct properties or attributes. This allows for flexibility in how spatial relationships are used during analysis. In contrast, the statistical approach integrates these relationships into mathematical formulas or adjusts the original dataset to account for them. This leads to a more holistic view of spatial data but can make the analysis more complex. The methods also differ in their focus. Statistical approaches tend to concentrate on intra-thematic relationships relationships within a single thematic layer of data. In contrast, learning-based approaches also consider inter-thematic relationships, or relationships between different thematic layers, which are essential for creating explanatory models. Lastly, methods like generalization and factorial analysis provide complementary tools. Generalization simplifies the data, making it easier to manage and interpret, while factorial analysis helps identify underlying correlations between variables. When used together, these methods enhance both the preparation and

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569| A Monthly Peer Reviewed & Referred Journal |

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1110116 |

interpretation of spatial data. For example, generalizing data before conducting classification or characterization allows for more meaningful and interpretable results.

V. CONCLUSION

This paper outlines various SDM methods from Statistics and Database research communities, comparing and classifying them to highlight their unique and combined advantages. Key future research areas include addressing temporal aspects of spatial data, optimizing spatial proximity relationships, and improving performance for large datasets. Enhancements in spatial indexes and pre-computation of neighborhood structures are essential for refining these techniques.

REFERENCES

1. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). "From data mining to knowledge discovery in databases." AI Magazine, 17(3), 37-54.
2. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers.
3. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann Publishers.
4. Miller, H. J., & Han, J. (Eds.). (2009). Geographic Data Mining and Knowledge Discovery. CRC Press.
5. Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). Geographic Information Systems and Science (4th ed.). Wiley.
6. Shekhar, S., & Chawla, S. (2003). Spatial Databases: A Tour. Prentice Hall.
7. Worboys, M. F., & Duckham, M. (2004). GIS: A Computing Perspective (2nd ed.). CRC Press.
8. Li, Z., Zhu, Q., & Gold, C. (2005). Digital Terrain Modeling: Principles and Methodology. CRC Press.
9. Albrecht, J. (2007). "Key Concepts and Techniques in GIS." SAGE Publications.
10. Hagenauer, J., & Helbich, M. (2012). "Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks." International Journal of Geographical Information Science, 26(6), 963-982.
11. Cheng, T., & Wang, J. (2009). "Integrated data mining and GIS modeling for simulating urban growth." International Journal of Geographical Information Science, 23(7), 937-960.
12. Brunsdon, C., & Comber, L. (2015). An Introduction to R for Spatial Analysis and Mapping. SAGE Publications.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details