REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-FEB-2002	Conference Proceedings, (refereed)	

4. TITLE AND SUBTITLE
Assessment Of Spatial Data Mining Tools For Integration Into An Object-Oriented GIS (GIDB)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0602435N

6. AUTHOR(S)
ROY VICTOR LADNER FREDERICK PETRY (Dr.)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Research Laboratory
Marine Geoscience Division
Stennis Space Center, MS 39529-5004

8. REPORTING ORGANIZATION REPORT NUMBER
NRL/PP/7440--02-1004

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release,distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A variety of data mining techniques are under evaluation on the spatial data of concern in our setting. We are planning to integrate a number of these techniques into our geospatial system (GIDB). Three approaches are under special consideration and are described in the paper. A COTS data mining system has been successfully used to develop predictive models of near-shore conditions such as wave height for naval amphibious operations. Attribute generalization was applied to seafloor data to obtain statements about conditions relevant to mine warfare. Finally an extension of association rule discovery applied to fuzzy spatial data that is under development is discussed.

15. SUBJECT TERMS
data mining techniques, spatial data, near-shore conditions

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Roy Ladner
unclassified	unclassified	unclassified	Unlimited	10	19b. TELEPHONE NUMBER *(Include area code)* 228-688-4679

Assessment of Spatial Data Mining Tools for Integration into an Object-oriented GIS (GIDB)

Roy Ladner[1] and Fredrick E. Petry[1]

[1]Naval Research Laboratory
Digital Mapping Charting and Geodesy Analysis Program
Stennis Space Center, MS 39529 USA
(rladner,fpetry)@nrlssc.navy.mil

Abstract. A variety of data mining techniques are under evaluation on the spatial data of concern in our setting. We are planning to integrate a number of these techniques into our geospatial system (GIDB). Three approaches are under special consideration and are described in the paper. A COTS data mining system has been successfully used to develop predictive models of near-shore conditions such as wave height for naval amphibious operations. Attribute generalization was applied to seafloor data to obtain statements about conditions relevant to mine warfare. Finally an extension of association rule discovery applied to fuzzy spatial data that is under development is discussed.

1 Introduction

Data mining or knowledge discovery generally refers to a variety of techniques that have developed in the fields of databases, machine learning and pattern recognition. The intent is to uncover useful patterns and associations from large databases. We are concerned with applications of data mining to spatial and temporal data.

In this paper we describe the geospatial system that will be the source of the data we are attempting to enhance with data mining for a number of Naval planning applications. Then we describe our experiences with three diverse data mining techniques that we have found applicable to the spatial data of interest in our setting and summarize their potential for integration into an overall enhanced system.

2 Background

We are developing approaches for spatial data mining in an environment in which there is considerable concern about the development of ways for processing large amounts of spatio-temporal data especially of oceanographic and littoral regions and including meteorological information. Our plan is to integrate the data mining techniques into the geospatial system described below. The ultimate goal is to provide

knowledge-enhanced information to decision tools that will be used by US Navy and Marine planners.

The Digital Mapping, Charting and Geodesy Analysis Program (DMAP) at the Naval Research Laboratory has been actively involved in the development of a digital geospatial mapping and analysis system since 1994. This work started with the Geospatial Information Database (GIDB™), an object-oriented, CORBA-compliant spatial database capable of storing multiple data types from multiple sources. Data is accessible over the Internet via a Java Applet [3].

The GIDB includes an object-oriented data model, an object-oriented database management system (OODBMS) and various analysis tools. While the model provides the design of classes and hierarchies, the OODBMS provides an effective means of control and management of objects on disk such as locking, transaction control, etc. The OODBMS in use is Ozone, an open-source database management system. This has been beneficial in several aspects. Among these, access to the source code allows customization and there are no costly commercial database licensing fees on deployment. Spatial and temporal analysis tools include query interaction, multimedia support and map symbology support. Users can query the database by area-of-interest, time-of-interest, distance and attribute. For example, statistics and data plots can be generated to reflect wave height for a given span of time at an ocean sensor. Interfaces are implemented to afford compatibility with Arc/Info, Oracle 8i, Matlab, and others.

The object-oriented approach has been beneficial in dealing with complex spatial data, and it has also permitted integration of a variety of raster and vector data products in a common database. Some of the raster data include satellite and motion imagery, Compressed ARC Digitized Raster Graphics (CADRG), Controlled Image Base (CIB), jpeg and video. Vector data includes Vector Product Format (VPF) products from the National Imagery and Mapping Agency (NIMA), Shape, real-time and in-situ sensor data and Digital Terrain Elevation Data (DTED). The VPF data includes such NIMA products as Digital Nautical Chart (DNC), Vector Map (VMAP), Urban Vector Map (UVMAP), Digital Topographic Data Mission Specific Data Sets (DTOP MSDS), and Tactical Oceanographic Data (TOD).

Over the years, the system has been expanded to include a communications gateway that enables users to obtain data from a variety of data providers distributed over the Internet in addition to the GIDB. These providers include Fleet Numerical Meteorology and Oceanography Center (FNMOC), USGS, Digital Earth/NASA, and the Geography Network/ESRI. A significant FNMOC product is the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) data. The atmospheric components of COAMPS are used operationally by the U.S. Navy for short-term numerical weather prediction for various regions around the world. Our communications gateway provides a convenient means for users to obtain COAMPS data and incorporate it with other vector and raster data in map form. The gateway establishes a well-defined interface that brings together such heterogeneous data for a common geo-referenced presentation to the user. An illustration of the interface for a typical data request is shown in Figure 1.

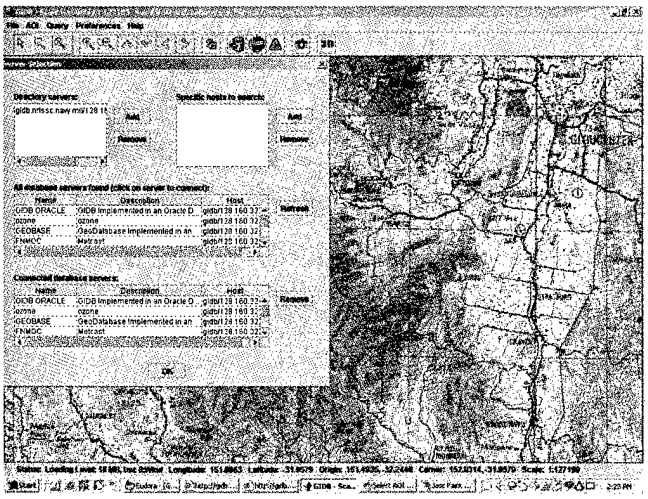

Fig. 1. The GIDB Interface.

3 Spatial Data Mining Techniques

3.1 Predictive Modeling

For this technique we used a COTS data mining system (the KXEN Knowledge Extraction Engines) based on the Support Vector (SV) approach. It is based on the VC (Vapnik-Chervonenkis) learning theory, which does not depend on dimensionality and can be applied to any function estimation problem [7].

The particular component we utilized is a regression algorithm, which builds predictive models. The support vector machine approach is an extension of the linear hyperplane classification of perceptrons to more complex surfaces. This is done by extending the measurement space so that it includes transformations of the raw variables. The distinct aspect of this approach is the score function called the margin. This is used to optimize the decision boundary between the classes such that it is likely to lead to the best possible generalization performance [5].

We have approximately 20 years of data observations of sea conditions at the Field Research Facility in Duck, North Carolina USA. Sensors record changing waves, winds, tides and currents on approximately an hourly basis. These are stored in the GIDB and were selectively used for the data mining experiments.

The application for which we wish to utilize this type of data is that of providing advisory information to tactical Naval planners for amphibious operations. One criti-

cal factor is the wave conditions near the beach for mine removal, landing craft operation, etc. In particular we considered wave height and wave periodicity. In doing this we were concerned with the ability to predict conditions that would jeopardize the mission. Thus we focus on prediction of the wave heights in the upper quartile of the recorded data. We needed to provide a qualified prediction as to whether the wave height would exceed a given operational capability for the equipment being planned for the mission as input by a user. For example based on prescribed operational requirements, the mission would have to be cancelled if the wave heights exceeded 1-½ meters.

For predicted wave heights we must consider two situations. The first is where we have overestimated wave heights. If a prediction is an overestimate that exceeds the specified limits, the operation might be erroneously scrubbed, thus missing an opportunity but not jeopardizing equipment or personnel.

On the other hand an underestimate might cause a planner to decide on continuing a mission in an unacceptable and potentially dangerous sea condition. This is clearly the more important error and the one presented in the table.

Training was done based on years 1999-2000 with a total of 13,591 observations. This model was then used to predict the wave heights for years 1993-1995 and 1997-1998. Data was not available for the entire year 1996. Consider the 1993 data in which there were 7773 observations (1943 in the upper quartile). In the upper quartile 1004 of these were underestimates (since very few are exact predictions in general we either have under or over estimates). The range of wave heights in the upper quartile was .60 - 2.0 meters. For the upper quartile, the average error for the underestimates was 24.7% and the average wave height of the underestimates was 0.95 meters. So the typical error in the underestimate was 0.27 meters in this critical range. This error was deemed to be generally acceptable although this is of course situation dependent.

Table 1. Duck, NC Wave Height Predictions

Year	Number of Observations	#Upper Quartile Under Estimate	Avg. Wave Height Under Estimate	Average % Under Estimate	Std. Dev
1993	7773	1004	0.95	24.75	0.19
1994	7735	625	0.92	24.86	0.16
1995	7767	876	0.94	29.79	0.20
1997	6162	1282	1.02	31.85	0.22
1998	6729	1461	1.06	32.10	0.22

3.2 Attribute Generalization

Both this technique of attribute-oriented induction and association rule generation are intended to provide a generalization or summarization of some potentially relevant aspects of the data being considered.

The attribute-oriented induction approach produces a generalized representation by either attribute removal or attribute generalization. After this step the processed data is

aggregated by merging identical tuples in the database and counting the number of tuples merged to indicate significance [4]. Attributes are removed if there is no hierarchy for the attribute or if it can be expressed in terms of higher-level concepts of other attributes. Attribute generalization examines an attribute to ascertain if there are too large a number of distinct values (exceeding a given threshold). Then if a generalization hierarchy is available for this attribute, it is generalized and the common tuples merged.

We applied this technique to sea bottom data from 10 locations (such as areas in the Philippines, Mediterranean, Persian Gulf, etc.). Here the intended application was to characterize various sea bottom areas for the planning of a mine deployment/hunting mission. The spatial data was queried to formulate the files from which the attribute generalization was done. The basic query was on bottom sediment classification as this was the major characteristic of interest to experts. The data consisted of polygons of the bottom types as classified. Depth was an estimate, and depth and area were binned into three categories

Shown below in Table 2 is generalization of data from the Onslow Bay area. The value of "any" is the root of the concept hierarchy to which the corresponding tuples have been generalized.

Table 2. Generalization of Bottom Data from the Onslow Bay Area.

Area	Type	Depth	Count
any	pure sand	shallow	46
any	pure sand	deep	44
any	sandy mix	deep	26
small	sandy mix	shallow	15
mid	sandy mix	shallow	9
mid	pure sand	mid	1
mid	sandy mix	mid	1
small	sandy mix	mid	1

3.3 Association Rules for Fuzzy Spatial Data

Association rules capture the idea of certain data items commonly occurring together. For example an analysis of the soils and vegetation in a certain region might reveal that 30% of the total area has co-occurring sandy soil and scrub cover and for any sandy soil area, 75% of these area had scrub cover. Thus we can obtain the rule

$$\textit{Sandy soil} \rightarrow \textit{Scrub cover} \qquad (1)$$

that could be used for planning and environmental decision makers. This rule is said to have a 75% degree of confidence and a 30% degree of support.

The process of generating rules requires the determination of the values of support and confidence and if a potential rule has values for these that exceed the user the user provided minimums it is called a strong rule [1]. Let $R = \{ T_1, T_2, \ldots.\}$ be the results of a query that obtains the data of interest. To determine if there is a strong relation

between values (possibly sets) A and B, the tuples in R must be examined and a count made of the number containing A and B where T_i contains A (B) if A (B) $\subseteq T_i$. Two measures are used to determine rules. First, the percentage of tuples that contain both A and B is called the support *s*. Second, if T_i contains A then it also contains B - called the confidence *c*. The Apriori algorithm [1] proceeds by first obtaining the sets of values (called itemsets) that satisfy the minimum support. It uses an iterative level-wise search where sets of k items are used to consider the set at the next level of k + 1 items. The final result is called the frequent itemsets. Then using the confidence value the strong association rules are extracted from the frequent itemsets

Now if the data we are interested in, as is typical of much spatial data [2], has uncertainty associated, we can model this using fuzzy sets [3]. So in general we will have fuzzy membership values associated to the tuples of R. The count to determine the support for finding frequent itemsets in the case of fuzzy data is developed by using the idea of the Σ count, which extends the ordinary concept of set cardinality to fuzzy sets [8]. Using this the fuzzy support count for the set A becomes:

$$FSC_R(A) = \Sigma \text{ Count } (A) = \sum_j \mu_{ti}. \tag{2}$$

Finally to produce the association rules from the set of relevant data R retrieved from the spatial database we can now extend the ideas of fuzzy support and confidence as

$$FS = FSC_R(A \cup B) / | R |, \tag{3}$$

$$FC = FSC_R(A \cup B) / FSC_R(A). \tag{4}$$

3.3.1 Example of Spatial Association Rules

We will consider an example that requires data mining on a spatial database to provide assistance in the logistical planning for a military operation. Assume that an area of operational interest has been divided into several zones (1, 2, etc.) and we would like to know some of the important relationships of relevant attributes in each zone to provide guidance in planning and selection of a zone for the particular mission. From this point of view small cities are of interest as they would have sufficient infrastructure but would not pose difficulties in which to operate, as would large cities. The major logistical concern is with transportation (railroads, highways, airfields) and terrain (soils, ground cover) within about 5 kilometers of the city.

The first step we must take to discovering rules that may be of interest in a given zone is to formulate a SQL query using the fuzzy function NEAR (Figure 2) to represent those objects within about 5 kilometers of the cities. Additionally we use the fuzzy function of Figure 3 to select the cities with a SMALL population.

Sample SQL.

```
SELECT City C, Road R, Railroad RR, Airstrip A,
Terrain T

FROM Area of Interest Zone 1

WHERE  {NEAR (C.loc, R.loc), NEAR (C.loc, RR.loc),

     NEAR (C.loc, A.loc), NEAR (C.loc, T.loc)}

and C.pop = SMALL

AT Threshold Levels = .80, .75, .70
```

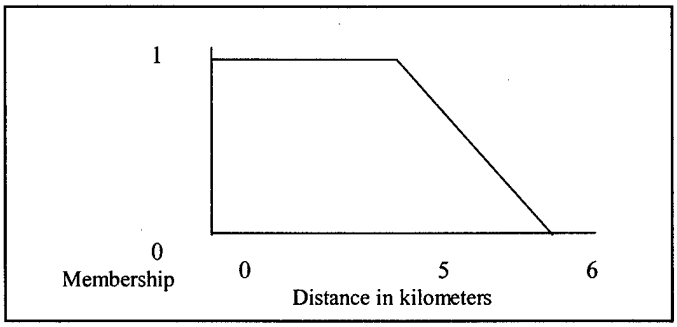

Fig. 2. Fuzzy Membership Function for Distance.

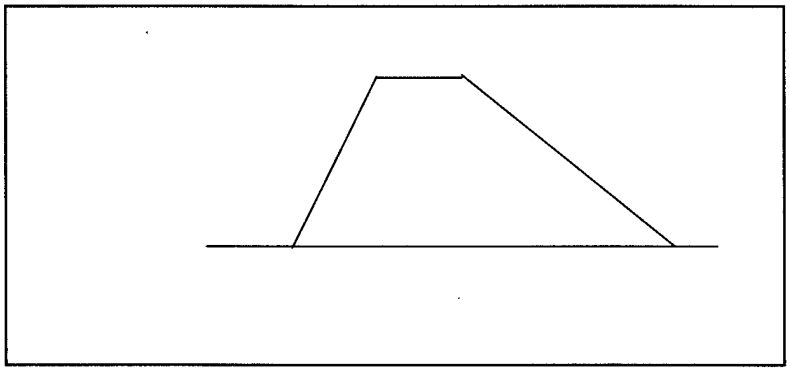

Fig. 3. Fuzzy Membership Function for Small City.

We evaluate for each city in a selected zone the locations of roads, railroads and airstrips using the NEAR fuzzy function. The terrain attribute value is produced by

evaluation of various factors such as average soil conditions (e.g. firm, marshy), relief (e.g. flat, hilly), coverage (fields, woods), etc. These subjective evaluations are then combined into one membership value, which is used to provide a linguistic label based on fuzzy functions for these. Note that the evaluation for terms such as "good" can be context dependent. For logistical purposes an open and flat terrain is suitable whereas for an infiltration operation a woody and hilly situation would be desirable.

Each attribute value in the intermediate relation then has a degree of membership. The three threshold levels in the query are specified for the NEAR, SMALL and the terrain memberships. The final relation R is formulated based on the thresholds and the tuple membership computed as a minimum of the individual memberships from the intermediate relation.

Table 3. The Final Result of the Example Query – R.

City	Roads	Railroads	Airstrips	Terrain	μ_t
A	Rte.10	RRx	None	Good	0.89
B	{Rte.5,Rte.10}	None	A2	Fair	0.79
F	Rte.6	RRx	None	Good	0.92
...	...	...	...	...	...

In R the value 'None' indicates for the attribute that no value was found NEAR – within the five kilometers. For such values no membership value is assigned and so μ_t is just based on the non-null attribute values in the particular tuple.

Now in the next step of data mining we generate the frequent itemsets from R using the fuzzy support count. At the first level for itemsets of size 1 (k=1), airstrips are not found since they do not occur often enough in R to yield a fuzzy support count above the minimum support count that was pre-specified. The level k=2 itemsets are generated from the frequent level 1 itemsets. Here only two of these possibilities exceed the minimum support and none above this level, i.e. k=3 or higher. This gives us the following table of frequent itemsets:

Table 4. The Frequent Itemsets Found for the Example.

k	Frequent Itemsets	Fuzzy Support-Count
1	{Road Near}	11.3
1	{Good Terrain}	10.5
1	{Railroad Near}	8.7
2	{Road Near, Good Terrain}	9.5
2	{Good Terrain, Railroad Near}	7.2

From this table of frequent itemsets we can extract various rules and their confidence. Rules will not be output unless they are strong – satisfy both minimum support and confidence. A rule produced from a frequent itemset satisfies minimum support by the manner in which frequent itemsets are generated, so it only necessary to use the fuzzy support counts from the table to compute the confidence. The small city clause that will appear in all extracted rules arises because this was the general condition that selected all of the tuples that appeared in query result R from which the frequent itemsets were generated.

Let us assume for this case that the minimum confidence specified was 85%. So, for example, one possible rule that can be extracted from the frequent itemsets in Table 4 is:

> If C is a small city and has good terrain nearby then there is a road nearby with 90% confidence.

Since the fuzzy support count for {Good Terrain} is 10.5 and the level 2 itemset {Road Near, Good Terrain} has a fuzzy support count of 9.5, the confidence for the rule is 9.5 / 10.5 or 90%. This is above the minimum confidence of 85%, so the rule is strong and will be an output of the data mining process.

If we had specified a lower minimum confidence such as 80% we could extract (among others) the rule:

> If C is a small city and has a railroad nearby then there is good terrain nearby with 83% confidence.

Since the fuzzy support count for {Railroad Near} and {Railroad Near, Good Terrain} are 8.7 and 7.2, the confidence is 7.2 / 8.7 or 83% and so this rule is also output.

4 Conclusions

We have found the data mining approaches we have described to be basically satisfactory in our preliminary evaluations. We are continuing with more extensive testing and evaluation with various sets of spatio-temporal data that are available to us.

There are three aspects of integrating these and similar data mining tools within our GIDB system. The first is the specification of the query to obtain the relevant data for the particular data mining tool(s) to be applied. As seen in some of our examples, for user query specification we should provide an interface that allows the query to be generated from selection of the data as it is displayed. Also as in the case of association rules, but applicable to other approaches as well, we must be able to deal with fuzzy function specification for the query. Considerable work has appeared that we can utilize for this in fuzzy set research [8]. Next for most techniques the user must provide some parameters and values for the techniques such as hierarchies in attribute generalization or support and confidence thresholds for association rule generation. Finally and related to the parameters specification is the issue of appropriate display of results to the user. This is more complex for the domain of spatial data. This is of concern not only for the final results but also based on the common assumption that the data mining will be an iterative process in which a user may be changing the input parameters and specifications as they obtain various preliminary results.

5 ACKNOWLEDGEMENTS

We would like to thank the Naval Research Laboratory's Base Program, Program Element No. 0602435N for sponsoring this research.

References

1. Agrawal R, Imielinski T, and Swami A, 1993, Mining Association Rules between sets of items in large databases. In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data.* New York, NY, ACM Press:207-216.
2. Burrough P, and Frank A (eds.), 1996 *Geographic Objects with Indeterminate Boundaries,* GISDATA Series Vol. 2, London, UK , Taylor and Francis.
3. Chung M, Wilson R, Ladner R, Lovitt T, Cobb M, Abdelguerfi M, and Shaw K, 2001 TheGeospatial Information Distribution System (GIDS). In Chaudhri A and Zicari R (eds) *Succeeding with Object Databases.* New York, NY, Wiley and Sons:357-378.
3. M. Cobb, F. Petry and V. Robinson (2000), Special Issue: Uncertainty in Geographic Information Systems and Spatial Data, *Fuzzy Sets and Systems*, 113, #1.
4. Han J, and Kamber M 2000 *Data Mining: Concepts and Techniques.* San Diego, CA Academic Press.
5. Hand D, Mannila H, and Smyth P, 2001, *Principles of Data Mining.* Cambridge, MA, MIT Press.
6. Lu W, Han J, and Ooi B, 1993, Discovery of general knowledge in large spatial databases. In *Proceedings of Far East Workshop Geographic Information Systems.* Singapore, World Scientific Press: 275-289.
7. Vapnik V, 1995, *The Nature of Statistical Learning Theory.* Berlin, GDR, Springer-Verlag.
8. Yen J, and Langari R, 1999, *Fuzzy Logic: Intelligence, Control and Information.* Upper Saddle River, NJ, Prentice Hall.