Transactions in GIS, 2003, 7(1): 123–138

Research Paper

Fuzzy Set Approaches to Spatial Data Mining of Association Rules

Roy Ladner
Naval Research Laboratory
Mapping, Charting and Geodesy
Stennis Space Center, Mississippi

Frederick E Petry
Naval Research Laboratory
Mapping, Charting and Geodesy
Stennis Space Center, Mississippi

Maria A Cobb
Department of Computer Science and Statistics
The University of Southern Mississippi

Abstract
This paper presents an approach to the discovery of association rules for fuzzy spatial data. Association rules provide information of value in assessing significant correlations that can be found in large databases. Here we are interested in correlations of spatially related data such as soil types, directional or geometric relationships, etc. We have combined and extended techniques developed in both spatial and fuzzy data mining in order to deal with the uncertainty found in typical spatial data.

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 shall describe a new approach that allows the discovery of association rules for fuzzy spatial data. Association rules provide information of value in assessing significant correlations that can be found in large databases. Here we are interested in correlations of spatially related data such as soil types, directional or geometric relationships, etc. We have combined and extended techniques developed in both spatial and fuzzy data mining to deal with the uncertainty found in typical spatial data.

Address for Correspondence: Fred Petry, Electrical Engineering and Computer Science Department, Tulane University, 310 Stanley Thomas Hall, New Orleans, LA 70118, USA Email: fep@eecs.tulane.edu

As an example consider using a spatial database to provide assistance in the logistical planning for a military operation. Then we might wish to uncover some of the important relationships of the data attributes in each area to provide guidance in the mission planning. We shall see in detail in the last section of the paper that one possible rule that might be discovered is of the form:

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

Such a rule incorporates fuzzy information in the linguistic terms used such as "small" and "nearby."

In this paper we first overview the background of data mining and uncertainty in spatial data (including the main research in spatial data mining). Then our approach to extracting spatial association rules for fuzzy data is described and an example is provided to illustrate the process.

2 Background

2.1 Data Mining

Although we are primarily interested here in some of the specific algorithms of knowledge discovery, we will first review the overall process of data mining. The initial steps are concerned with preparation of data, including data cleaning intended to resolve errors and missing data and integration of data from multiple heterogeneous sources. Next are the steps needed to prepare for actual data mining. These include the selection of the specific data relevant to the task and the transformation of this data into a format required by the data mining approach. These steps are sometimes considered to be those in the development of a data warehouse, i.e., an organized format of data available for various data mining tools. There are a wide variety of specific knowledge discovery algorithms that have been developed (Han and Kamber 2000). These discover patterns that can then be evaluated based on some interestingness measure used to prune the huge number of available patterns. Finally as true for any decision aid system, an effective user interface with visualization/alternative representations must be developed for the presentation of the discovered knowledge.

Specific data mining algorithms can be considered as belonging to two categories – descriptive and predictive data mining. In the descriptive category are class description, association rules and classification. Class description can either provide a characterization or generalization of the data or comparisons between data classes to provide class discriminations. Association rules are the main focus of this paper and correspond to correlations among the data items. They are often expressed in rule form showing attribute-value conditions that commonly occur at the same time in some set of data. An association rule of the form $X \rightarrow Y$ can be interpreted as meaning that the tuples in the database that satisfy the condition X also are "likely" to satisfy Y, so that the "likely" implies this is not a functional dependency in the formal database sense. Finally, a classification approach analyzes the training data (data whose class membership is known) and constructs a model for each class based on the features in the data. Commonly, the outputs generated are decision trees or sets of classification rules. These can be used both for the characterization of the classes of existing data and to allow the classification of data in the future, and so can also be considered predictive.

Predictive analysis is also a very developed area of data mining. One very common approach is clustering. Clustering analysis identifies the collections of data objects that are similar to each other. The similarity metric is often a distance function given by experts or appropriate users. A good clustering method produces high quality clusters to yield low inter-cluster similarity and high intra-cluster similarity. Prediction techniques are used to predict possible missing data values or distributions of values of some attributes in a set of objects. First, one must find the set of attributes relevant to the attribute of interest and then predict a distribution of values based on the set of data similar to the selected objects. There are a large variety of techniques used, including regression analysis, correlation analysis, genetic algorithms and neural networks.

Finally, a particular case of predictive analysis is time-series analysis. This technique considers a large set of time-based data to discover regularities and interesting characteristics. One can search for similar sequences or subsequences, then mine sequential patterns, periodicities, trends and deviations.

2.2 Uncertainty in Spatial Data

The need to handle imprecise and uncertain information concerning spatial data has been widely recognized (Goodchild 1990), particularly in the field of geographical information systems (GIS). The value of GIS as a decision-making tool is dependent on the ability of decision-makers to evaluate the reliability of the information on which their decisions are based. Users of GIS technology must therefore be able to assess the nature and degree of error in spatial databases, track this error through GIS operations and estimate accuracy for both tabular and graphic output products. There are a variety of aspects of potential errors in GIS encompassed by the general term "accuracy." However, here we are only interested in those aspects that lend themselves to modeling by fuzzy set techniques.

Many operations are applied to spatial data under the assumption that features, attributes and their relationships have been specified a priori in a precise and exact manner. However, inexactness often exists in the positions of features and the assignment of attribute values and may be introduced at various stages of data compilation and database development. Models of uncertainty have been proposed for spatial information that incorporate ideas from natural language processing, the value of information concept, non-monotonic logic and fuzzy sets, and evidential and probability theory (Stoms 1987).

A number of researchers in the GIS and spatial database area have recently considered models of spatial data using fuzzy set approaches, as in the modeling of geographic objects with indeterminate boundaries (Burrough and Frank 1996) and the Journal of Fuzzy Sets and Systems recently published a special issue on uncertainty in GIS and spatial data (Cobb et al. 2000). Some early work by geographers in the 1970s utilized fuzzy sets in topics such as behavioral geography and geographical decision-making (Gale 1972, Pipkin 1978, Leung 1979). However, the first consistent approach to the use of fuzzy set theory as it could be applied in GIS was developed by Robinson (Robinson and Frank 1985, Robinson 1988, 1990). More recently, there have been a number of efforts utilizing fuzzy sets for spatial databases including some for capturing spatial relationships (Cobb and Petry 1998, Guesgen and Albrecht 2000), querying spatial information (Morris and Petry 1998, Wang 2000), and object-oriented modeling (George et al. 1992, Morris et al. 1998, Cross and Firat 2000).

2.3 Fuzzy Data Mining

An early and continuing significant application of fuzzy sets has been in pattern recognition, especially fuzzy clustering algorithms (Bezdek 1974). Hence, much of the effort in fuzzy data mining has been by the use of fuzzy clustering and fuzzy set approaches in neural networks and genetic algorithms (Hirota and Pedrycz 1999). In fuzzy set theory an important consideration is the treatment of data from a linguistic viewpoint. From this an approach has developed that uses linguistically quantified propositions to summarize the content of a database, by providing a general characterization of the analyzed data (Yager 1991, Kacprzyk and Zadrozny 2000). A common organization of data for data mining is the multidimensional data cube in data warehouse structures. Treating the data cube as a fuzzy object has provided another approach for knowledge discovery (Laurent 2002).

Fuzzy data mining for generating association rules has been considered by a number of researchers. There are approaches using the SETM (Set-oriented mining) algorithm (Shu et al. 2001) and other techniques (Bosc and Pivert 2001) but most have been based on the important Apriori algorithm. Extensions have included fuzzy set approaches to quantitative data (Zhang 1999, Kuok et al. 1998), hierarchies or taxonomies (Chen et al. 2000, Lee 2001), weighted rules (Gyenesei 2000) and interestingness measures (de Graaf et al. 2001, Gyenesei 2001). For our work here, we focus on a basic extension of the Apriori algorithm motivated by some of the above developments, but we also envision future work that includes the use of fuzzy taxonomies and fuzzy weights for discovering spatial association rules.

3 Spatial Data Mining

There is considerable interest in spatial data mining, but only a few major research efforts have been developed in this area. A major difference between data mining in ordinary relational databases and in spatial databases is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhoods (such as topological, distance and direction relations), which are used by spatial data mining algorithms.

A very active and influential data mining research group is that led by Han in Vancouver and presently at Illinois. They have investigated several approaches to spatial data mining and have developed a system called GeoMiner (Han et al. 1997) based on these techniques. One approach developed a generalization-based knowledge discovery mechanism that integrated attribute-oriented induction on non-spatial data and spatial merge and generalization on the spatial data (Lu et al. 1993). The CLARANS clustering algorithm is a randomized search for an optimal cluster. Another spatial data mining approach was based on CLARANS and produced high-level non-spatial description of objects in every cluster using attribute-oriented induction (Ng and Han 1994).

For this research, the most important work is the development of an approach for mining strong association rules in geographic information databases (Koperski and Han 1995). This approach uses an SQL-like spatial data mining query interface as developed for GeoMiner. This provides the subset of the spatial database over which the rule discovery is performed. From this subset the spatial predicates of interest such as

intersect, adjacent, etc. are then explicitly materialized. The Apriori algorithm (to be discussed in detail in the next section) (Agrawal et al. 1993) is applied over these data to extract the association rules. If there is a concept hierarchy for the data and/or the spatial predicates, a multi-level approach to the Apriori algorithm allows rules to be extracted at any desired level. Another interesting approach uses a hierarchy of the topological relations of objects with a broad boundary (Clementini and DiFelice 1997) to mine spatial association rules for objects with uncertainty (Clementini et al. 2000).

A research group in Munich (Ester et al. 2000) has developed a set of database primitives for mining in spatial databases that are sufficient to express most of the algorithms for spatial data mining and that can be efficiently supported by a DBMS. They have found that the use of such database primitives enables the integration of spatial data mining with existing DBMSs and speeds-up the development of new spatial data mining algorithms. The database primitives are based on the concepts of neighborhood graphs and neighborhood paths. Effective filters allow restriction of the search to such neighborhood paths "leading away" from a starting object. Neighborhood indices materialize certain neighborhood graphs to support efficient processing of the database primitives by a DBMS. For spatial characterization it seems important that class membership of a database object is not only determined by its non-spatial attributes but also by the attributes of objects in its neighborhood. In spatial trend analysis, patterns of change of some non-spatial attributes in the neighborhood of a database object were determined. Spatial trends can be thought of as describing the regular change of non-spatial attributes when moving away from certain start objects for which both global and local trends can be distinguished.

Another approach has been taken by the use of spatial autocorrelation rather than materializing spatial predicates. The system is used to predict locations using map similarity (Chawla et al. 2000). It has four components – map similarity measures, parametric functions for spatial models, a discretized parameter space and the search algorithm. The search explores the parameter space to discover the parameter value tuple maximizing the map similarity measure. These parameter values thus indicate the parametric function to use as the possible spatial model.

4 Fuzzy Spatial Data Mining

4.1 Introduction

In this section we will describe the combination of certain approaches we have surveyed in the past two sections concerning fuzzy data mining and spatial data mining. To date no efforts have appeared in the literature that have specifically investigated fuzzy spatial data mining. The setting for which we are developing our approaches for fuzzy spatial data mining is an environment in which considerable concern about uncertainty in several respects arises. The objective is to develop ways of processing large amounts of spatio-temporal data especially of oceanographic and littoral regions and including meteorological information. We are studying spatial data mining in several aspects including the fuzzy techniques in this paper. 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 (Chung et al. 2001).

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 time the system has been expanded to include a communications gateway enabling 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 Plate 1, See plate section.

4.2 Spatial Data Uncertainty Classes

To investigate the potential complications caused by uncertainty in discovering spatial association rules in spatial databases, we have to consider the sorts of data that will occur. There are two particular classes of spatial data that we shall focus upon which we call: type I – spatial or spatially related attributes and type II – spatial relationships.

Type I. For this category we may have "spatially related" attributes such as soil types: {sandy, loamy, . . . }. These classes are of course not crisp and here we will use a

similarity table approach (Buckles and Petry 1982) to matching values as was done in previous work on the conflation of VPF data (Cobb et al. 1998). It is also possible to allow numeric as well as scalar values. For example, a fuzzy number such as "about 3 cm" of rainfall at a location could be used. A similarity approach to the matching of fuzzy numbers that has been previously developed could then be used here (Buckles and Petry 1984).

Type II. The major types of spatial relationships that can be considered are geometrical orientation and topological relationships. In particular, we will consider relationships such as distance and direction with fuzzy membership functions with topological relationships to be considered for future work. In either of these cases we may have an ontological concept hierarchy, which may also involve uncertainty relationships in the hierarchy that can be modeled with fuzzy sets. Such a hierarchy can be used to produce multi-level association rules, which have a finer gradation of meaning.

In the data mining process there are three particular aspects to be considered for issues of uncertainty: (1) querying in which the spatially relevant data is obtained for the data mining; (2) the actual Apriori algorithm which produces the association rules; and (3) relevance evaluation procedures that assess and prune the association rules produced. In this paper we focus only on the first two and leave the last as a topic of future research.

4.3 Fuzzy Querying for Spatial Data Mining

First, we examine the management of uncertainty in the data-mining query, which develops the table on which the Apriori algorithm is applied. We will follow the basic approach taken in GeoMiner (Koperski and Han 1995, Han et al. 1997) and extend it as necessary. One problem is the form of the resultant relation obtained from the querying involving the types of spatial data. Since uncertainty can arise from the similarity degree of matching of type I data in the query and/or spatial predicates (type II), we are faced with the choice of maintaining uncertainty measures for the individual attributes or formulating an overall uncertainty/membership value for the resultant tuples in the relation. A number of approaches have been taken to this concern in the fuzzy database literature (Petry 1996), but here our concern is determining the form that will be effective for data mining.

A crucial aspect of the query for the formulation of the data over which the data mining algorithm will operate is the selection of a spatial predicate that identifies the specific spatial region or area of interest (AOI). This is closely related to the property that causes objects of interest to cluster in space, which is the so-called first law of geography: "Everything is related to everything else but nearby things are more related than distant things" (Tobler 1979). A common choice for this is some distance metric such as a NEAR predicate; however, other spatial predicates such as CONTAINS or INTERSECTS could also be used.

Let us consider an SQL form of a query:

SELECT Attributes A, B
FROM Relation X
WHERE
(X.A = α and NEAR (X.B, β))
AT Threshold Levels = M, N

Table 1 The intermediate result of the SQL form query (R_{int} =)

A	B
$<a_1, \mu_{a1}>$	$<b_1, \mu_{b1}>$
$<a_2, \mu_{a2}>$	$<b_2, \mu_{b2}>$
$<a_3, \mu_{a3}>$	$<b_3, \mu_{b3}>$
. . .	. . .
$<a_i, \mu_{ai}>$	$<b_i, \mu_{bi}>$

To match the values in the query, we have for the attribute A, a spatially related attribute, a similarity table of its domain values, and for B, a spatial attribute such as location, the NEAR predicate can be evaluated. Since these values may not be exact matches, the intermediate resultant relation R_{int} will have to maintain the degree of matching. The final query results are chosen based on the values specified in the threshold clause. Results for attribute *A* are based on the first value *M* and similarly those for B are based on *N*. The level values are typically user specified as linguistic terms that correspond to such values. The intermediate step of the query evaluation is shown in Table 1.

In the table a_i is a value of *A* from *X*, and μ_{a_i} is the similarity of a_i and α

$$\mu_{a_i} = \text{sim}(a_i, \alpha)$$

For example, let the domain be soil types and if a_i = loam and α = peat then the similarity might be

$$\text{sim(loam, peat)} = 0.75$$

and if the threshold level value *N* in the query were lower than 0.75, we could possibly retain this in the final relation as the value and membership

$$<\text{loam}, 0.75>$$

Similarly for the location attribute, where $b_i = (13.1, 74.5)$ and $\beta = (12.9, 74.1)$, we might have

$$<(13.1, 74.5), 0.78>$$

if the coordinates (in some measure) are "near" by a chosen fuzzy distance measure

$$\mu_{near} = \text{NEAR}((13.1, 74.5), (12.9, 74.1)) = 0.78.$$

Figure 1 shows a typical fuzzy NEAR function that might be used here to represent: "within a distance of about 5 kilometers or less."

However rather than retaining individual memberships, the data mining algorithm will be simplified if we formulate a combined membership value. Thus the final result R is obtained by evaluating each tuple relative to the threshold values and assigning a tuple membership value based on the individual attribute memberships combined by the commonly used min operator. So the tuple $<a_i, b_i, \mu_t>$ will appear in the final result relation R if and only if

$$(\mu_{a_i} > M) \text{ and } (\mu_{b_i} > N)$$

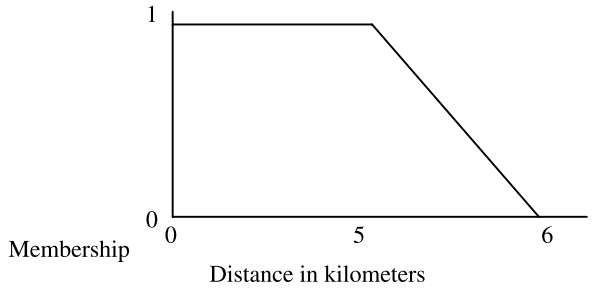

Figure 1 Fuzzy Membership Function for Distance

Table 2 The final result of the SQL form query using threshold values (R)

A	B	μ_t
a_1	b_1	Min(μ_{a1}, μ_{b1})
a_3	b_3	Min(μ_{a3}, μ_{b3})
. . .	. . .	
a_i	b_i	Min(μ_{ai}, μ_{bi})

and the tuple membership is

$$\mu_t = \min(\mu_{a_i}, \mu_{b_i})$$

If the rows (tuples) 1 and 3 from R_{int} are such that the memberships for both columns are above their respective thresholds, then these are retained. However for tuple 2, let it be the case that $\mu_{a2} > M$, but the second attribute's membership is $\mu_{b2} < N$. Then this tuple will not appear in the final result relation R as shown in Table 2.

4.4 Association Rules for Spatial Data

4.4.1 Association rules

Association rules capture the idea of certain data items commonly occurring together and have often been considered in the analysis of a "marketbasket" of purchases. For example, a delicatessen retailer might analyze the previous year's sales and observe that of all purchases 30% were of both cheese and crackers and, for any of the sales that included cheese, 75% also included crackers. Then it is possible to conclude a rule of the form:

$$\text{Cheese} \rightarrow \text{Crackers}$$

This rule is said to have a 75% degree of confidence and a 30% degree of support. A retailer could use such rules to aid in the decision process about issues such as placement of items in the store, marketing options such as advertisements and discounts and so forth. In a spatial data context an analysis of various aspects of a certain region might produce a rule associating soils and vegetation such as:

$$\text{Sandy soil} \rightarrow \text{Scrub cover}$$

that could be used for planning and environmental decision-making.

This particular form of data mining is largely based on the Apriori algorithm developed by Agrawal et al. (1993). Let a database of possible data items be:

$$D = \{d_1, d_2, \ldots d_n\}$$

and the relevant set of transactions (sales, query results, etc.):

$$R = \{T_1, T_2, \ldots\}$$

where $T_i \subseteq D$. We are interested in discovering if there is a relationship between two sets of items (called itemsets) X_j, X_k; X_j, $X_k \subseteq D$. For such a relationship to be determined, the entire set of transactions in R must be examined and a count made of the number of transactions containing these sets, where a transaction T_i contains X_m if $X_m \subseteq T_i$. This count, called the support count of X_m, $SC_R(X_m)$, will be appropriately modified for fuzzy sets.

There are then two measures used in determining rules: the percentage of T_i's in R that:

1. contain both X_j and X_k (i.e. $X_j \cup X_k$) – called the support s
2. if T_i contains X_j then T_i also contains X_k – called the confidence c.

The support and confidence can be interpreted as probabilities:

1. s – Prob $(X_j \cup X_k)$ and
2. c – Prob $(X_k \mid X_j)$

We assume the system user has provided minimum values for these in order to generate only sufficiently interesting rules. A rule whose support and confidence exceeds these minimums is called a strong rule.

The overall process for finding strong association rules can be organized as a three step process:

1. Determine the frequent itemsets – this is most commonly accomplished with variations of the Apriori algorithm,
2. Extract strong association rules from the frequent itemsets, and
3. Assess generated rules with interestingness measures.

The first step is to compute the frequent itemsets F which are the subsets of items from D, such as $\{d2, d4, d5\}$. The support count SC_R of each such subset must be computed and the frequent itemsets are then only those whose support count exceeds the minimum support count specified. This is just the product of the minimum support specified and the number of transactions or tuples in R. For a large database this generation of all frequent itemsets can be very computationally expensive. The Apriori algorithm is an influential algorithm that makes this more computationally feasible. It basically uses an iterative level-wise search where sets of k items are used to consider sets at the next level of $k + 1$ items. The Apriori property is used to prune the search as seen in the discussion below.

After the first and more complex step of determining the frequent itemsets, the strong association rules can easily be generated. The first step is to enumerate all subsets of each frequent itemset F: $f_1, f_2, \ldots f_i \ldots$ Then for each f_i, calculate the ratio of the support count of F and f_i, i.e.

$$SC_R(F)/SC_R(f_i)$$

Note that all subsets of a frequent itemset are frequent (Apriori property) and so the support counts of each subset will have been computed in the process of finding all frequent itemsets. This greatly reduces the amount of computation needed.

If this ratio is greater that the minimum confidence specified then we can output the rule:

$$f_i \rightarrow \{F - f_i\}$$

The set of rules generated may then be further pruned by a number of correlation and heuristic measures.

4.4.2 Fuzzy spatial association rules

We are now prepared to consider how to extend approaches to generating association rules to process the form of the fuzzy data we have developed for the spatial data query. A number of previous approaches were discussed in section 2.3 and our extension is based primarily on Chen et al. (2000). Recall that in order to generate frequent itemsets, we must count the number of transactions T_j that support an itemset X_j. In the ordinary Apriori algorithm one simply counts the occurrence of a value as 1 if in the set, or if not in the set – 0. Here, since we have obtained from the query a membership degree for the values in the transaction, we must modify the support count SC_R. To achieve this we will use the Σ Count operator that extends the ordinary concept of set cardinality to fuzzy sets (Yen and Langari 1999). Let A be a fuzzy set, then the cardinality of A is obtained by summation of the membership values of the elements of A:

$$Card(A) = \sum Count(A_i) = \sum \mu_A(y_i);\ \ y_i \in A$$

Using this the fuzzy support count for the set X_j becomes:

$$FSC_R(X_j) = \sum Count(X_j) = \sum_i \mu_{T_i};\ \ X_j \subseteq T_i$$

Note that membership of T_i is included in the count only all of the values of the itemset X_j are included in the transaction, i.e. it is a subset of the transaction.

Finally to produce the association rules from the set of relevant data R retrieved from the spatial database, we will provide our extension to deal with the resulting frequent itemsets. For the purposes of generating a rule such as $X_j \rightarrow X_k$ we can now extend the ideas of fuzzy support and confidence as:

$$FS = FSC_R(X_j \cup X_k)/|\ R\ |$$

$$FC = FSC_R(X_j \cup X_k)/FSC_R(X_j)$$

4.4.3 Example

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 then take to discovering rules that may be of interest in a given zone is to formulate the SQL query as we have described above using the fuzzy

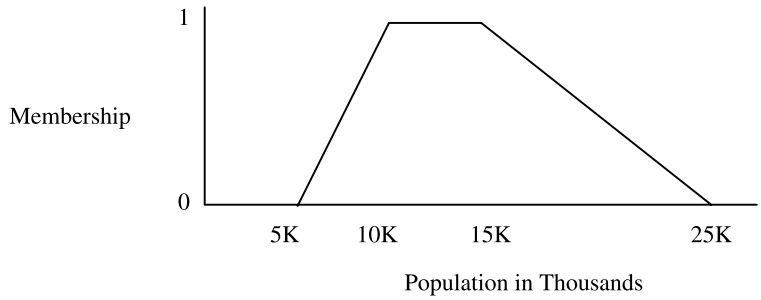

Figure 2 Fuzzy Membership Function for Small City

function NEAR (Figure 1) to represent those objects within about 5 kilometers of the cities. We use the fuzzy function of Figure 2 to select the cities with a small population.

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

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 that 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 previously described (see Table 3 for details).

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 table of frequent itemsets reproduced in Table 4.

From this table of frequent itemsets we can extract various rules and their confidence. Rules will not be output unless they are strong – in other words, they must 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

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
...	...	...	...	...	...

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

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%. Since this is above the minimum confidence of 85%, this 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.

5 Conclusions

We have presented an approach to the discovery of association rules for fuzzy spatial data where we are interested in correlations of spatially related data such as soil types,

directional or geometric relationships, etc. We have combined and extended techniques developed in both spatial and fuzzy data mining in order to deal with the uncertainty found in typical spatial data. We plan to also extend our effort to topological relationships (Krishnapuram et al. 1993, Zhan 1998, Bloch 1999).

Some areas for future work include hierarchies, weights and interestingness measures. The type II data of spatial relationships have cases that involve hierarchies. For example, the NEAR predicate could be organized as a hierarchy of relationships such as contains, intersects, etc. Each of these might then have a different strength in the hierarchy as well as being defined by fuzzy membership functions. The combination of these values is complex and must be worked out in the knowledge discovery context. Another case is a directional relationship such as NORTH that could be a hierarchy of northeast, north, northwest for which north would have the highest strength in the hierarchy.

Consider from our example in the previous section that airstrips did occur as a frequent itemset because of the low number of airstrips occurring in the spatial data. However, rules involving airstrips might be considered as of great importance to a particular mission. We would like to provide a fuzzy weighting to such items of importance so that they might exceed the minimum support and thus appear in final rules without having to lower minimum support and create a plethora of rules involving other objects. In general it is possible that even with a high minimum support and confidence a large number of rules can be generated. Measures of interest are being considered to either prune or prioritize the final rules that are output.

We are in the process of testing this approach on near-shoreline data that are of concern to the Naval operational environment. We are also examining the extension of the attribute-oriented generalization on spatial data approach (Lu et al. 1993) to fuzzy spatial data. Our overall goal is to be able to integrate the data mining tools and techniques that prove useful for spatial data into our GIDB geospatial system.

Acknowledgements

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

References

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: 207–16

Bezdek J 1974 Cluster validity with fuzzy sets. *Journal of Cybernetics* 3: 58–72

Bloch I 1999 Fuzzy relative position between objects in image processing: A morphological approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21: 657–64

Bosc P and Pivert O 2001 On some fuzzy extensions of association rules. In *Proceedings of IFSA-NAFIPS 2001*. Piscataway, NJ, IEEE Press: 1104–9

Buckles B and Petry F 1982 A fuzzy representation for relational databases. *International Journal of Fuzzy Sets and Systems* 7: 213–26

Buckles B and Petry F 1984 Extending the fuzzy database with fuzzy numbers. *Information Sciences* 34: 45–55

Burrough P and Frank A (eds) 1996 *Geographic Objects with Indeterminate Boundaries*. London, Taylor and Francis

Chawla S, Shekar S, Wu W, and Ozesmi U 2000 Predicting locations using map similarity (Plums): A framework for spatial data mining. In *Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, ACM Press: 243–51

Chen G, Wei Q, and Kerre E 2000 Fuzzy data mining: Discovery of fuzzy generalized association rules. In Bordogna G and Pasi G (eds) *Recent Issues on Fuzzy Databases*. Heidelberg, Physica-Verlag: 45–66

Chung M, Wilson R, Cobb M A, Petry F E, and Shaw K 2001 Querying multiple data sources via an object-oriented spatial query interface and framework. *Journal of Visual Languages and Computing* 12: 37–60

Chung M, Wilson R, Ladner R, Lovitt T, Cobb M A, Abdelguerfi M, and Shaw K 2001 The Geospatial Information Distribution System (GIDS). In Chaudhri A and Zicari R (eds) *Succeeding with Object Databases*. New York, John Wiley and Sons: 357–78

Clementini E and DeFelice P 1997 Approximate topological relations. *International Journal of Approximate Reasoning* 16: 173–204

Clementini E, DeFelice P, and Koperski K 2000 Mining multiple-level spatial association rules for objects with a broad boundary. *Data and Knowledge Engineering* 34: 251–70

Cobb M A and Petry F E 1998 Modeling spatial data within a fuzzy framework. *Journal of the American Society for Information Science* 49: 253–66

Cobb M A, Chung M, Foley H, Petry F E, and Shaw K 1998 A rule-based approach for the conflation of attributed vector data. *GeoInformatica* 2: 1–29

Cobb M A, Petry F E, and Robinson V B (eds) 2000 Special issue: Uncertainty in geographical information systems and spatial data. *International Journal of Fuzzy Sets and Systems* 113: 1–159

Cross V and Firat A 2000 Fuzzy objects for geographical information systems. *International Journal of Fuzzy Sets and Systems* 113: 19–36

de Graaf J, Kosters W, and Witteman J 2001 Interesting fuzzy association rules in quantitative databases. In *Principles of Data Mining and Knowledge Discovery – LNAI 2168*. Berlin, Springer-Verlag: 140–51

Ester M, Fromelt A, Kriegel H, and Sander J 2000 Spatial data mining: Database primitives, algorithms and efficient DBMS support. *Data Mining and Knowledge Discovery* 4: 89–125

Gale S 1972 Inexactness, fuzzy sets and the foundation of behavioral geography. *Geographical Analysis* 4: 337–49

George R, Buckles B, Petry F, and Yazici A 1992 Uncertainty modeling in object-oriented geographical information systems. In *Proceedings of 1992 Conference on Database and Expert System Applications*. Berlin, Springer-Verlag: 77–86

Goodchild M and Gopal S (eds) 1990 *The Accuracy of Spatial Databases*. London, Taylor and Francis

Guesgen H and Albrecht J 2000 Imprecise reasoning in geographic information systems. *International Journal of Fuzzy Sets and Systems* 113: 121–31

Gyenesei A 2000 Mining Weighted Association Rules for Fuzzy Quantitative Items. Turku, Turku Center for Computer Science Technical Report No 346

Gyenesei A 2001 Interestingness measures for fuzzy association rules. In *Principles of Data Mining and Knowledge Discovery – LNAI 2168*. Berlin, Springer-Verlag: 152–64

Han J and Kamber M 2000 *Data Mining: Concepts and Techniques*. San Diego, CA, Academic Press

Han J, Koperski K, and Stefanovic N 1997 GeoMiner: A system prototype for spatial data mining. In *Proceedings of the 1997 ACM-SIGMOD International Conference on Management of Data*. New York, ACM Press: 553–6

Hirota K and Pedrycz W 1999 Fuzzy computing for data mining. In *Proceedings of the IEEE* 87: 1575–99

Kacprzyk J and Zadrozny S 2000 On combining intelligent querying and data mining using fuzzy logic concepts. In Bordogna G and Pasi G (eds) *Recent Issues on Fuzzy Databases*. Heidelberg, Physica-Verlag: 67–81

Krishnapuram R, Keller J, and Ma Y 1993 Quantitative analysis of properties and spatial relations of fuzzy image regions. *IEEE Transactions on Fuzzy Systems* 1: 222–33

Koperski K and Han J 1995 Discovery of spatial association rules in geographic information databases. In *Proceedings of the Fourth International Symposium on Large Spatial Databases*. Berlin, Springer-Verlag: 47–66

Kuok C, Fu A, and Wong H 1998 Mining fuzzy association rules in databases. *ACM SIGMOD Record* 27: 41–6

Laurent A 2003 Fuzzy multidimensional databases for fuzzy-OLAP mining. *IEEE Transactions on Fuzzy Systems* 11: in press

Lee K 2001 Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies. In *Proceedings of IFSA-NAFIPS 2001*. Piscataway, NJ, IEEE Press: 2977–82

Leung Y 1979 Locational choice: A fuzzy set approach. *Geographic Bulletin* 19: 28–34

Lu W, Han J, and Ooi B 1993 Discovery of general knowledge in large spatial databases. In *Proceedings of the Far East Workshop Geographic Information Systems*. Singapore, World Scientific Press: 275–89

Morris A and Petry F E 1998 Design of fuzzy querying in object-oriented spatial data and GIS. In *Proceedings of NAFIPS 98*. Piscataway, NJ, IEEE Press: 211–5

Morris A, Petry F E, and Cobb M A 1998 Fuzzy object-oriented database modeling of spatial data. In *Proceedings of the IPMU Conference*. Paris, EDK Press: 604–11

Ng R and Han J 1994 Efficient and effective clustering method for spatial data mining. In *Proceedings of 1994 International Conference on Very Large Databases*. San Francisco, CA, Morgan Kaufmann: 144–55

Petry F E 1996 *Fuzzy Databases: Principles and Applications*. Norwell, MA, Kluwer

Pipkin J 1978 Fuzzy sets and spatial choice. *Annals of the Association of American Geographers* 68: 196–204

Robinson V B and Frank A 1985 About different kinds of uncertainty in geographic information systems. In *Proceedings of the AUTOCARTO 7 Conference*. Falls Church, VA, American Society for Photogrammetry and Remote Sensing: 440–9

Robinson V B 1988 Implications of fuzzy set theory for geographic databases. *Computers, Environment and Urban Systems* 12: 89–98

Robinson V B 1990 Interactive machine acquisition of a fuzzy spatial relation. *Computers and Geosciences* 6: 857–72

Shu J, Tsang E, and Yeung D 2001 Query fuzzy association rules in relational databases. In *Proceedings of IFSA-NAFIPS 2001*. Piscataway, NJ, IEEE Press: 2989–93

Stoms D 1987 Reasoning with uncertainty in intelligent geographic information systems. In *Proceedings of GIS'87*. Falls Church, VA, American Society for Photogrammetry and Remote Sensing: 693–9

Tobler W 1979 Cellular geography. In Gale S and Olsson G (eds) *Philosophy in Geography*. Dortrecht, Riedel: 379–86

Wang F 2000 A fuzzy grammar and possibility theory-based natural language user interface for spatial queries. *International Journal of Fuzzy Sets and Systems* 113: 147–59

Yager R 1991 On linguistic summaries of data. In Piatesky-Shapiro G and Frawley (eds) *Knowledge Discovery in Databases*. Boston, MA, MIT Press: 347–63

Yen J and Langari R 1999 *Fuzzy Logic: Intelligence, Control and Information*. Upper Saddle River, NJ, Prentice Hall

Zhan F 1998 Approximate analysis of topological relations between geographic regions with indeterminate boundaries. *Soft Computing* 2: 28–34

Zhang W 1999 Mining fuzzy quantitative association rules. In *Proceedings of IEEE International Conference on Tools with Artificial Intelligence*. Piscataway, NJ, IEEE Press: 99–102