MINING FUZZY SPATIAL ASSOCIATION RULES FROM IMAGE DATA

By

George Brannon Smith

A Thesis Proposal
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Computer Science
in the Department of Computer of Science

Mississippi State, Mississippi

August 2000

MINING FUZZY SPATIAL ASSOCIATION RULES FROM IMAGE DATA

By

George Brannon Smith

Approved:

Susan M. Bridges Associate Professor of Computer Science (Major Professor)	Julia Hodges Professor/Department Head of Computer Science (Committee Member)
Hasan Jamil Assistant Professor of Computer Science (Committee Member)	A. Wayne Bennett Dean of the College of Engineering

Name: George Brannon Smith

Date of Degree: August 18, 2000

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: Dr. Susan Bridges

Title of Study: MINING FUZZY SPATIAL ASSOCIATION RULES FROM IMAGE DATA

Pages in Study: 46

Candidate for Degree of Master of Science

The work of Agrawal et al. on the type of data mining known as association rule mining has been the basis for continuous research over the past seven years. Kuok, Fu and Wong have extended association rules with the fuzzy set theory of Zadeh to build a system more adapted to real world data. Meanwhile Bloch has recently applied this same fuzzy set theory to the area of image analysis, particularly to the task of finding relationships between image objects. Koperski and Han have been the leaders in adapting general data mining techniques, including association rules, to the domain of spatial data. This proposed thesis offers a synthesis of these techniques to form a unified system for generalized image analysis, specifically finding general fuzzy rules about the relationships of objects in image data sets.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

With the rapid advancement in computing power, storage capacity, and collection techniques, more and more spatial image data are being collected. This data explosion is compounded by the fact that because the data sources, namely the earth, sea and atmosphere, are dynamic, data often have to be recollected to be of real use, creating another dimension to the data. But the raw data itself is only the first step in understanding. Analysis is required to find useful generalized information, but considering the volume of the data, human expert analysis becomes a tedious and tiresome job. This problem has given rise to many algorithms and systems seeking to extract useful generalized observations of various types. Such analysis commonly goes by the broad term *Knowledge Discovery in Databases* (KDD) or sometimes simply *Knowledge Discovery* - the finding of implicit and previously unknown or unseen information in a database by the application of computer algorithms. In fact, Fayyad et al. declare that "Knowledge Discovery is the most desirable end-product of computing" (Fayyad et al. 1996). Unfortunately, they also concede that it is " ... one of the most difficult computing challenges to do well"(Fayyad et al. 1996). Doing it well is a pursuit in which many researchers are engaged and they have turned out several techniques towards that end. A useful system could be constructed by combining some of these existing techniques into a multi-layer system.

1.1 Motivation

As mentioned before, much data is collected from spatial sources such as remote sensing of the oceans and terrain. We are often interested in the nature of the many components of such images, objects and regions (e.g., a stand of trees, coral reefs, etc.) and may apply KDD techniques such as classification to extract these. But now that we know the classes of the various entities in an image or more likely a large set of images, what next? These components do not exist by themselves; they exist in an image (and on the earth in the case of real world remote sensing) along with many other objects. We would like to know the relationships among the objects. For example, "Sandy ocean bottom #235 occurs 2 miles to the right of coral reef #769" in an image (which may in turn mean east or down current in the real world). Or "Grassy meadow area #114 is located 5 miles above deciduous forest #80." However, considering the volume of data, even individual spatial relations may not be of interest to us. Rather, generalized information on aggregates of the data is more immediately useful. How does one class of region relate to other classes? An answer, courtesy of a specific KDD system, might be "Lakes (a class) tend to occur next to mountains (another class)" or "Sandy bottoms (a class) tend to occur close to and below coral formations (another class)." A more recognizably useful scenario might be:

- Discovered "Rocky bottoms usually occur to the west (having corrected for real world orientation) of bottom type #12".
- Bottom type #12 has often proven a good place to drill for oil.
- Remote sensing data shows no bottom type #12 - only a rocky bottom.
- However, knowing the general relationship of the two classes, let's focus on the area west of the rocky bottom and see what turns up.

A real world example is the joint project between the *Naval Oceanographic Office* (NAVO) and the *Mississippi State University Computer Science Department*, in which undersea side-scan SONAR data is being analyzed by a research group. The image data is being processed by region growing techniques and the Bayesian classifier Autoclass, which partitions the sea-floor data into discrete regions of relatively few classes. Doing this effectively and efficiently is the current focus of the research. Generalizing these results in their spatial context is a next step. A rule induction system could be used to take the classified image set and discover the generalized spatial nature of the component parts, thereby providing more human usable information about the data.

1.2 Background

In 1993, Agrawal, Imielinski and Swami presented their theory of *association rule mining* for finding generalizations in a database, with their initial test domain being retail store transaction data. The application is not limited to this domain, but can be applied to general boolean attribute data (e.g., item is/is not purchased, person is/is not married, system is/is not operational). Indeed subsequent research has extended the technique to categorical attributes (e.g., car make in {Ford, GM, Toyota, Daimler, etc.}) and quantitative attributes (e.g., client Age in [18-24], [25-30], [30-35], etc.). Kuok, Fu and Wong have even extended that with fuzzy set theory to extract fuzzy association rules from quantitative data (e.g., if client Age in 'Young Adult' to some degree then income is 'Low' to a certain degree for some percentage of the database).

Among the researchers exploring association rules are Koperski and Han who have been applying a variety of other data mining techniques to spatial data, data

pertaining to the position of objects in a space. Koperski and Han attempt to make generalizations along the lines of: 65% of houses are west_of a lake. While they do acknowledge the mining of image or raster data, they typically work on data actually in a spatial database system in which the data has already been broken down and analyzed to a certain degree for its initial insertion into such a system.

Somewhat similar is the work of Bloch, who has developed techniques for finding spatial relationships between objects. The chief differences are:

1. She is working mainly on raw image data (such as MRI scans) rather than prepared spatial database entries.
2. She is using fuzzy set theory to take into account the varying morphology (i.e., size and shape) of the objects in question, which can have an impact on how best to describe the spatial relationship.

The result is not a single precise number but a range, which may seem like a failing, but which actually captures the imprecision of the real relationship.

The aforementioned fuzzy set theory is a generalization of classical set theory that was first proposed by Zadeh in 1965. Whereas an element of a classical or *crisp* set is either totally in a given set or not in it at all, fuzzy set theory allows for elements to be in a set to a degree ranging from total inclusion to total exclusion. The former works fine for domains where set membership is indeed a binary relationship such as belonging to a sports team or being a US State. The latter is useful for situations such as the set of all TALL people or OLD people, sets which have no sharp boundaries unless they are arbitrarily imposed. The applications above exploit this feature for their own designs. An object A is to the RIGHT of object B to some degree in a Bloch image. A person is in the 'Young Adult' set to some degree: 1.0 for a 22 year old; 0.3 for a 16 year old.

1.3 Hypothesis and Main Goals

Fuzzy spatial techniques based on the work of Bloch (1999) can be applied to a partitioned and classified image data source, possibly NAVO sea-floor data, to generate intermediate data about the images and the objects therein. Adaptations of standard association rule mining algorithms can then be applied to this meta-data to generate classic association rules or even fuzzy association rules describing the nature of the image components. The synthesis of these elements will provide generalized information useful for our purposes.

1.4 Organization

The remaining chapters are organized as follows:

- Literature Review: A survey of the source work used for this research.
- Research Plan:
  - Methodology: A statement of the formal methods on which the research will be based including techniques and algorithms from the fields of spatial relations, data/association rule mining, and fuzzy sets; also some of the pitfalls, and possible solutions, that may be encountered along the way.
  - Hypothesis: A restatment of the hypothesis.
  - Plan of Attack: An outline of how the aforementioned techniques will be implemented and integrated to form an actual system to be used for the experiments.
  - Publication Plan: A list of journals and conferences where finished paper may be submitted.

CHAPTER II

LITERATURE REVIEW

2.1 Fuzzy Logic/Fuzzy Sets

In 1965, Lotfi Zadeh presented his controversial theory of fuzzy sets(Zadeh 1965), and he is generally recognized as the founder of the discipline, though Black (1937) did discuss the related concept of vagueness as far back as 1937, and some have pointed out that there were considerations on the issue of a region between true and false as far back as Plato (Brule 1985). What Zadeh proposed was a generalization of classical set theory to handle the inexactness, approximation and what he calls *elasticity* of real life, such as the concepts *tallness* or *nearness*.

We will assume the reader has an understanding of the basics of classical set theory, its properties, and operations such as union and intersection, and so will not review those fundamentals here except to say a word about the notion of a *characteristic function*. One of the common ways of describing a set is by giving its characteristic function(Schmucker 1984), and in fact a set and its characteristic function are often used interchangeably. A characteristic function is one that maps the elements of the universe of discourse onto the two element set $\{0, 1\}$, sometimes called a *valuation set*(Dubois and Prade 1980). This means that given a universe X and a subset A:

$$char_A(x) = \begin{cases} 1 & \text{iff } x \in A \\ 0 & \text{iff } x \notin A \end{cases}$$

For example, given a universe $X = \{a, b, c, d, e, f, g\}$, a characteristic function could yield a subset A:

$$A = \{a/1, b/0, c/0, d/1, e/0, f/1, g/0\}$$

as a result. Customarily, false members are not shown, nor are the characteristic values (there could be only one) resulting in the more familiar roster form

$$A = \{a, d, f\}$$

(which we can do for a finite subset of a finite universe). This works fine for domains in which set membership is truly binary such as the set of all students currently enrolled at Mississippi State University, or the set of all Nobel Laureates. It does not work for less precise everyday concepts such as the set of all TALL men in the room or the set of all stores CLOSE to a residence.

To handle these imprecise scenarios, Zadeh proposes defining set membership not on the finite two-valued set $\{0, 1\}$ but rather over the real interval $[0, 1]$. Set elements can still take the old characteristic values or *membership values* of 0 and 1 - but they can also now take on say 0.7 or 0.34 indicating partial set membership or a degree of membership. This fuzzy set can be described by a different kind of characteristic function, a *membership function* denoted by $\mu_A(x)$ and taking values in $[0, 1]$.

For example, given the same universe as above $X = \{a, b, c, d, e, f, g\}$, a membership function could yield a fuzzy subset A:

$$A = \{a/0.9, b/0.23, c/0, d/0.7, e/0.4, f/1, g/0\}$$

as a result. Again, completely false members (i.e., with a 0 degree of membership) are not shown. However, other fuzzy membership values are retained. A roster form in this case would be

$$A = \{a/0.9, b/0.23, d/0.7, e/0.4, f/1\}$$

This concept of partial membership allows us to represent real world concepts more in line with the everyday manner in which we consider them. Returning to the example of the set of TALL people, classifying 7'1" Frank as TALL ($\mu_{TALL}(Frank) = 1$) and 4'10" Cathy as not TALL ($\mu_{TALL}(Cathy) = 0$) was not a problem before with classical sets. But what about 6'2" David? The typical natural language response is probably along the lines of "sort of" or "somewhat" and assigning classical set membership can be difficult and unsatisfactory. However, $\mu_{TALL}(David) = 0.7$ better captures the imprecision of the real world description. Along these same lines, the fuzzy membership function eliminates the crisp boundary between classical set inclusion and exclusion which is often capricious and arbitrary. For example the classical set RICH may be designated by a characteristic function by which all people with income at or above $80,000 are members. Someone with income of $79,999 is not RICH while someone making $80,000 is RICH. So the difference between RICH and not RICH is now a mere dollar which is unrealistic. A gradual fuzzy membership function remedies this: Warren Buffett is still 100% RICH and someone barely scraping by is still 0% RICH but the path from the one to the other is smoother, with the $79K employee being in RICH to some degree.

Exactly what the fuzzy membership function looks like is not carved in stone but is chosen by the user based on his needs and domain knowledge. There are however several common standard membership functions including trapezoidal, triangular, Gaussian, bell curves, and sigmoid (Yen and Langari 1998). Some examples are (Yen and Langari 1998):

$$triangle(x : a, b, c) = \begin{cases} 0 & \text{iff } x < a \\ (x-a)/(b-a) & \text{iff } a \leq x \leq b \\ (c-x)/(c-b) & \text{iff } b \leq x \leq c \\ 0 & \text{iff } x > c \end{cases}$$

$$gauss(x : m, \sigma) = exp\left(-\frac{(x-m)^2}{\sigma^2}\right).$$

Again, which is used depends on the domain, and the values of the a, b, c, m, σ may depend on domain specific knowledge gathered by the user or through some other means, or whatever other information he wants.

Since fuzzy set theory is a generalization of classical set theory, it stands to reason that there are operations analogous to those on classical sets, including, but not limited to, complement/negation, intersection/disjunction, union/conjunction, and cardinality.

The complement of a fuzzy set is simply 1 - membership. That is:

$$\mu_{\neg A}(x) = 1 - \mu_A(x).$$

This makes sense; if an item is in set Q to degree 0.75, it stands to reason that it is NOT in set Q to degree 0.25.

Things are more difficult with union and intersection. Fuzzy set theory does not have an intersection/disjunction operation per se but rather a class of operations called triangular norms or *t-norms*. A t-norm is a two-valued function that maps pairs $[0,1] \times [0,1]$ onto a single value in $[0,1]$. A t-norm is formally defined as satisfying a certain set of axioms.

A t-norm, denoted $t(a,b)$, satisfies the following axioms for any $a,b,c,d \in [0,1]$, (where a,b,c,d could be interpreted as fuzzy membership values $\mu_A(x)$, $\mu_B(x)$, $\mu_C(x)$, and $\mu_D(x)$ respectively, where $x \in X$, the universe):

1. $t(0,0) = 0$; $t(a,1) = t(1,a) = a$ (function boundaries)
2. $t(a,b) \leq t(c,d)$ whenever $a \leq c$ and $b \leq d$ (monotonicity)
3. $t(a,b) = t(b,a)$ (commutativity)
4. $t(a,t(b,c)) = t(t(a,b),c)$ (associativity)

Also, the the inequality

$$t_w(a,b) \leq t(a,b) \leq min(a,b)$$

where

$$t_w(a,b) = \begin{cases} \min(a,b) & \text{if } max(x,y) = 1 \\ 0 & x,y < 1 \end{cases}$$

must be satisfied (Yen and Langari 1998).

In a similar fashion, fuzzy union/conjunction is implemented by a class of operations called t-conorms. For the same conditions above, a t-conorm, denoted $s(a, b)$, satisfies the following:

1. $s(1, 1) = 1$; $s(a, 0) = s(0, a) = a$ (function boundaries)
2. $s(a, b) \leq s(c, d)$ whenever $a \leq c$ and $b \leq d$ (monotonicity)
3. $s(a, b) = s(b, a)$ (commutativity)
4. $s(a, t(b, c)) = s(s(a, b), c)$ (associativity)

Also,the inequality

$$max(a, b) \leq s_w(a, b) \leq s_w(a, b)$$

where

$$s_w(a, b) = \begin{cases} \max(a,b) & \text{if } min(x, y) = 0 \\ 1 & x, y > 0 \end{cases}$$

must be satisfied.(Yen and Langari 1998) Furthermore, these operations are pairwise related:

$$t_m(a, b) = 1 - s_m(1 - a, 1 - b)$$

and vice versa.

Given these operator classes, one can then say:

$$\mu_{A \cap B}(x) = t_m(\mu_A(x), \mu_B(x))$$

where t_m is some t-norm.

Since, these are classes of operations, we can choose the ones that best suit our domain, provided they adhere to the axioms and are implemented in dual pairs. For example, we may choose to use the Einstein product t-norm out of the huge number of choices (Zimmerman 1996):

$$t_{Einstein}(a,b) = \frac{a \cdot b}{2 - [a + b - a \cdot b]}.$$

However, when the time comes to apply the fuzzy set union operation, the corresponding Einstein sum t-conorm should be used for consistency.

That said, Bellman and Giertz (1973) have made a case for the *standard* fuzzy set union and intersection operations first used by Zadeh, as the best and most natural extensions of classical set theory:

$$\begin{aligned} \mu_{A \cap B}(x) = &\quad t_{std}(\mu_A(x), \mu_B(x)) = \quad min(\mu_A(x), \mu_B(x)), \text{ and} \\ \mu_{A \cup B}(x) = &\quad s_{std}(\mu_A(x), \mu_B(x)) = \quad max(\mu_A(x), \mu_B(x)). \end{aligned}$$

Another often useful operation or property of classical sets is *cardinality* which is basically a count of the number of elements in a set, assuming the set in question is finite. However, since elements of fuzzy sets are often not completely in the set, it seems incorrect to count each member regardless of degree of membership. Fortunately the answer is quite simply the sum of the fuzzy memberships (Dubois and Prade 1980):

$$|A| = \sum_{x \in X} \mu_A(x).$$

This formula also works correctly on a classical set.

As useful as these analogous operations are, it is important to point out what is noticeably absent in fuzzy sets. Returning once again to the earlier example, $X = \{a, b, c, d, e, f, g\}$, with a fuzzy subset

$$A = \{a/0.9, b/0.23, c/0, d/0.7, e/0.4, f/1, g/0\}$$

using the aforementioned complement operation yields:

$$\neg A = \{a/0.1, b/0.77, c/1, d/0.3, e/0.6, f/0, g/1\}.$$

Applying the standard fuzzy union and intersection operations results in:

$$\begin{aligned} A \cup \neg A &= \{a/0.9, b/0.77, c/1, d/0.7, e/0.6, f/1, g/1\}, \text{ and} \\ A \cap \neg A &= \{a/0.1, b/0.23, c/0, d/0.3, e/0.4, f/0, g/0\}. \end{aligned}$$

Recall from classical set theory that given a boolean set B in a universe X:

$$\begin{aligned} B \cup \neg B &= X, \text{ and} \\ B \cap \neg B &= \emptyset \end{aligned}$$

These are known as the the *Law of Excluded Middle* and the *Law of Contradiction* (Yen and Langari 1998) respectively, and they are conspicuously absent from generalized fuzzy theory. For some, this represents a failing of fuzzy logic, but for others it is thoroughly in line with expectations. The very nature of fuzzy set membership is that an element can be both in a set and in its negation *simultaneously.*

2.2 Fuzzy Relative Positioning

Dr. Isabelle Bloch of the Ecole Nationale Supérieure des Télécommunications in Paris has been applying the theory of fuzzy sets to the problem of analysis of spatial relations of objects in images (Bloch 1999). We often want to describe the relationships between objects with imprecise natural language terms like "above" or "to the right of." This is made difficult by the facts that

- in real life objects are simply not always *exactly* "to the left of" other objects but can be offset.
- the shape and size, the morphology, of objects has an impact on their spatial relationship.

If a man is standing to the exact left of a woman and then shuffles forward a few inches, he does not completely cease to be to her left - just to a lesser degree. A man standing in the angle of a large L -shaped building may be both to the right and below it (assuming a bird's-eye view). As he walks away, he may continue to be to the right and be below for several yards.

Bloch offers a new methodology for dealing with this phenomenon in such domains as medical imaging, in both 3D and 2D space. Bloch actually utilizes fuzzy theory in two different ways. The main one is the directional relationship between objects in an image space as mentioned above. In other words, object A can be to the right of object B to some fuzzy degree. But objects themselves can also be defined as fuzzy sets to account for possible spatial imprecision, defined on the universe $\mathcal{S}$, the image itself. The elements of such a set are simply the points or pixels in the object. Being a fuzzy set, an object A can be represented by fuzzy membership function $\mu_A(x), x \in \mathcal{S}$ on the interval $[0, 1]$. It should be noted that

since crisp sets are specializations of fuzzy sets, this definition can handle crisp objects taking only the values 0 and 1 as well.

Since Bloch's method is designed to work in 3D space, a direction is defined by two angles $\alpha_1 \in [0, 2\pi]$, a heading, and $\alpha_2 \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, an elevation where $\alpha = (\alpha_1, \alpha_2)$. This then yields the unit vector

$$\vec{u}_{\alpha_1,\alpha_2} = (\cos\alpha_2 \cos\alpha_1, \cos\alpha_2 \sin\alpha_1, \sin\alpha_2)^t$$

for evaluation. If we choose to work only in 2D space, then $\alpha_2 = 0$ or we could simply let $\alpha = \alpha_1$.

Given this α, we want to find the degree to which an object A, defined by $\mu_A(x)$, is in that direction of a reference object R ($\mu_R(x)$), which is simply another object in the image universe. This is done by first constructing what Bloch refers to as a *landscape* which is the area

> ... around the reference object R ... such that the membership value of each point corresponds to the degree of satisfaction of the spatial relation under examination(Bloch 1999).

This landscape is itself a fuzzy set, denoted by $\mu_\alpha(R)$, where "in the direction α" is the spatial relation in question. A visualization of such a landscape is shown in Figure 2.2, generated from the sample image in Figure 2.1. Each point in the image space, excluding the points of the square object itself, takes a grey value from black to white indicating the degree of its membership in the fuzzy relation "in the direction $\alpha = 0$ of the square reference object, R". White indicates full membership, black indicates null membership, and the greys are various degrees in between. The object itself has been explicitly colored black.

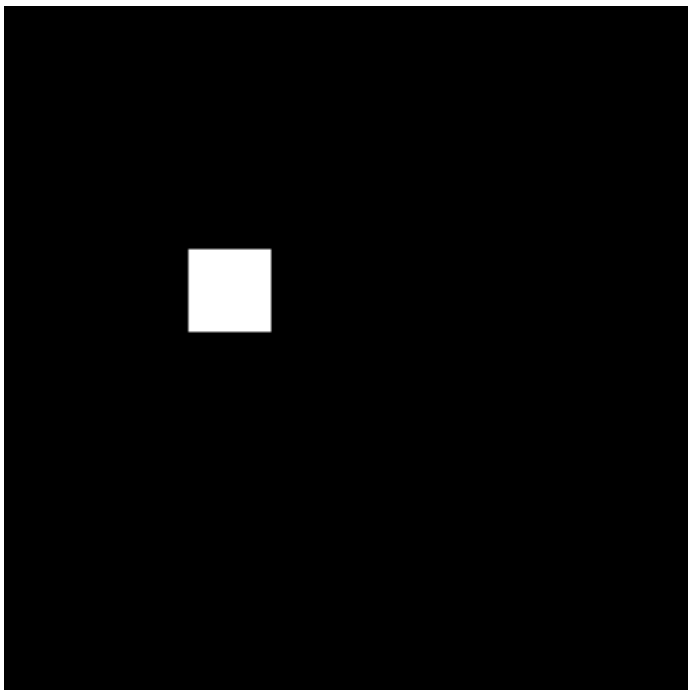

Figure 2.1: A shape in an image space

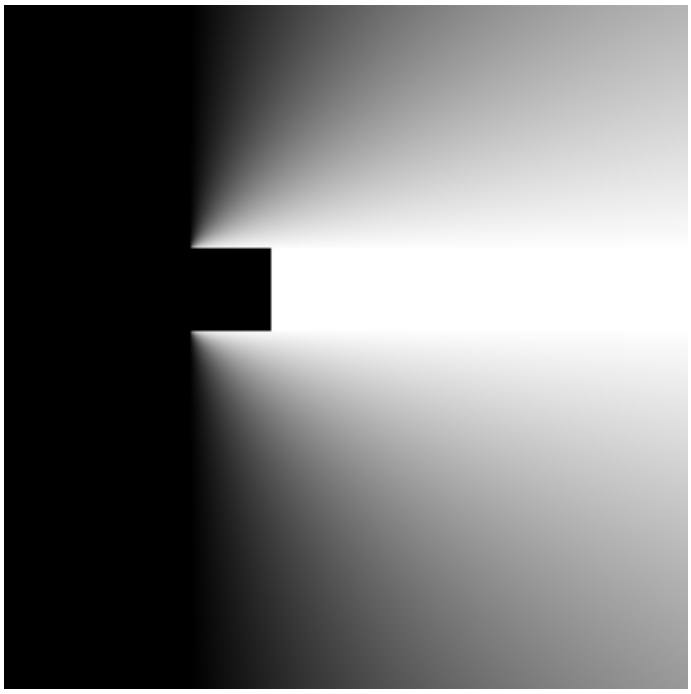

Figure 2.2: Landscape for $\alpha = 0$

Theoretically this landscape potentially extends over the entire universe, but practically speaking a full roster generation is neither computationally desirable nor necessary. We are only interested in the overlap between the landscape of R and object A, specifically a function of $\mu_\alpha(R)(x)$ and $\mu_A(x)$, which will indicate the relationship between the points in A and the points in R.

The final result is not actually a single fuzzy value, but rather three making an interval and a likely estimate. The values of the above function iterated over the points in the objects are evaluated in the following forms:

Possibility

$$\prod_{\alpha_1,\alpha_2}^{R}(A) = \sup_{x\in\mathcal{S}} t[\mu_\alpha(R)(x), \mu_A(x)]$$

Necessity

$$N^R_{\alpha_1,\alpha_2}(A) = \inf_{x\in\mathcal{S}} s[\mu_\alpha(R)(x), 1-\mu_A(x)]$$

Mean

$$M^R_{\alpha_1,\alpha_2}(A) = \frac{1}{|A|}\sum_{x\in\mathcal{S}} \mu_A(x)\mu_\alpha(R)(x)$$

where $t[]$ is a fuzzy t-norm, $s[]$ is a fuzzy t-conorm, supremum is least upper bound, and infinum is greatest lower bound. The *possibility* is an optimistic result - a kind of max; *necessity* is a pessimistic result - a kind of min; *mean* is the average membership degree over all the points. Together they express the value and interval $M \in [N, \Pi]$. In Bloch's opinion this can further represent the imprecision of the spatial relationship.

As stated in the section on fuzzy sets, the membership function for a fuzzy set is chosen by the user based on the various requirements of the domain. For the

landscape function $\mu_\alpha(R)$ Bloch has chosen a basic linear function

$$\mu_\alpha(R)(P) = \max\left(0, 1 - \frac{2\beta_{min}(P)}{\pi}\right), P \in \mathcal{S}$$

which turns the angle β into a fuzzy value in $[0, 1]$ (Again, while $P \in \mathcal{S}$, computationally this is typically only calculated when $P \in A$, A being the other object in question). $\beta_{min}(P)$ is the minimum of all the angles $\beta(P, Q)$ where $Q \in R$, and

$$\beta(P, Q) = \arccos\left[\frac{\overrightarrow{QP} \cdot \vec{u}_{\alpha_1,\alpha_2}}{\|\overrightarrow{QP}\|}\right]$$

where $\overrightarrow{QP}$ is the vector from the point in R to the point in A. So we are essentially iterating over the points in A and finding the best angle to a point in R. This angle is compared to the specified angle α.

2.3 Association Rules

The idea of *association rule mining* as a useful KDD technique was first put forth in 1993 by Agrawal, Imielinski and Swami (1993) based on experiments with retail store transaction database testbeds.

In short, an association rule takes the form

$$antecedent \Rightarrow consequent(c\%confidence, s\%support)$$

where the antecedent consists of one or more items in the set of transactions being mined and the consequent consists of only one item not in the antecedent.

The *support* is the percentage of all transactions in the database containing both the antecedent and the consequent, and it measures the significance of the rule in the database. The *confidence* is the percentage of those transactions containing the antecedent which also contain the consequent and measures the strength of the rule. Together these two are the principal *constraints* on the association rule mining process.

A classic example of an association rule is:

$$beer \Rightarrow chips(87\%\ confidence, 3\%\ support)$$

Here 3% of all transactions (i.e., possibly a basket of goods purchased by a customer on a store visit) in the database contain both beer and chips. It should be noted that the opposite does not necessarily follow; transactions containing chips may not include beer as well. Store management may be able to use rules such as these to make decisions about store operations: where to place displays, when to have sales and on what products, etc.

Following the notation of Agrawal, Imielinski and Swami, we formally define association rules as follows:

$\mathcal{I}$ is the set of all items aka binary attributes in the database; $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$
T is the transaction database
t is a transaction; $t \subseteq \mathcal{I}$
X is a set of some of the items; $X \subseteq \mathcal{I}$
t satisfies X if, for each item in X, t contains that item; $X \subseteq t$
An association rule is $X \Rightarrow I_j$ *s.t.* $X \subseteq \mathcal{I} \wedge I_j \in \mathcal{I} \wedge I_j \notin X$
$X \Rightarrow I_j$ is satisfied in T with confidence $0 \leq c \leq 1$ iff at least c% of transactions in T which satisfy X also satisfy I_j.
The result is sometimes written $X \Rightarrow I_j \mid c$

Items are sometimes called binary attributes because they can be view from a boolean point view. Either I_n is in a transaction or it is not. Both transactions and itemsets are simply sets of items in the universe $\mathcal{I}$.

Additional constraints can be placed on antecedent and consequent to generate more immediately useful rules such as

"Antecedent *must* contain I_n" or

"Consequent *must* be an element of a given set Q".

The application of user-defined constraints other than support and confidence is a subject for study in itself (Ng et al. 1998, Silberschatz and Tuzhilin 1996, Smith 1999).

Support is an important constraint on the resultant rule, measuring statistical significance and often business significance (given that these initial experiments are retail oriented). Low support may indicate low preferences or rules that are not worth further examination.

Agrawal et al. (1993) decompose the Association Rule Mining problem into two main subproblems:

1. Finding *large* itemsets (aka *frequent* itemsets). These are itemsets that exceed the minimum support threshold supplied by the user. These large itemsets will be used to generate the rules. They may be further pruned by additional constraints. Not surprisingly, itemsets that do not meet the threshold are *small* or *infrequent* itemsets.

2. Generating all the rules for each previously found large itemset. Formally:

   large itemset $Y = \{I_1, I_2 \ldots, I_k\}, k \geq 2$ will generate rules $X \Rightarrow I_j$ such that $X \subseteq Y$, $\mid X \mid = k-1$ and $I_j = Y - X$, i.e., $X \Rightarrow Y - X$. Put another way, Y is the union of the antecedent and consequent of every rule derived from it.

   To enforce the user specified minimum confidence (*minconf*) threshold c:

   if $support(Y)/support(X) > c$ then the rule satisfies c.

   Obviously we must have the support of X available for this. NOTE: As for minsupp, we know that Y is large from (1), and we know that X is large because *all subsets of a large itemset are also large.*

Agrawal, Imielinski and Swami (1993) mainly focus on the first subproblem and offer a custom algorithm (later know as AIS), bearing in mind that

- measuring all possible itemsets on one pass could result in checking 2^m combos in a database where the number of items, m, could easily reach 1000; and
- measuring only size k itemsets on k^{th} pass, building up candidates on each could result in many passes (k = 1000).

In a subsequent paper, Agrawal and Srikant (1994) offer some refinements, particularly:

- A relaxation of the requirement that rule consequents be a single item - sets of items are now allowed.
- A new algorithm, Apriori, and its variants, which are significantly faster than the previous AIS algorithm.

Apriori and its brethren tackle the first subproblem previously laid out - that of finding all large itemsets from which actual rules would then be derived. It attempts to relieve problems of the AIS algorithm in which too many candidates that turn out small are generated. Simply put, Apriori takes the set of large $k-1$ itemsets (i.e., knowledge it has *a priori*) joins them to create a new set of k-itemsets, and prunes out the ones whose subsets are small. This follows from the logic that since any subset of a large set must be large then:

a) large k-1 itemsets can alone be used to to generate candidate large k-itemsets.

b) Any k-itemset with small $k-1$ subsets *cannot* be large.

Apriori is significant because it is efficient, relatively simple, and it scales well and has thus been used by much subsequent research as either a baseline for comparison of new techniques or as the core algorithm for the development of new techniques and domain specific modifications.

These initial forays into the realm of association rules dealt primarily with binary attributes making the results *Boolean Association Rules*. Either a transaction contains "soda" or it does not - true or false. But other kinds of data exist, specifically quantitative such as price, which have values over a range, and categorical such as brand (e.g., Dell, IBM, Compaq, etc.). Srikant and Agrawal (1996) want to mine *Quantitative Association Rules* in order to gain information about this kind of data and would like to do it with the rule infrastructure already in place. A simple approach would be to partition the ranges and categories thereby mapping them onto binary attributes, allowing us to apply the foundational binary association rule techniques discussed earlier (e.g., Apriori algorithm). For example

- $Price = 1200 / Price \neq 1200$

- $Brand = Dell / Brand \neq Dell$

or if there are many values, use intervals:

- $Price \in [1000, 1500] / Price \notin [1000, 1500]$.

A transaction in this database would now contain (or not contain) the item *Price = 1200*, which could be mined like other items. Unfortunately, this partitioning can actually cause some contradictory problems. A few large intervals can encompass many tuples and thus have high support while many smaller intervals, while being more selective, may yield low and perhaps unacceptable support. On the other hand, a small interval, say $Age \in [21, 22]$ may provide for a nice specific high confidence rule such as

$$Age \in [21, 22] \rightarrow GraduatedCollege(80\%)$$

whereas widening the interval to gain support brings in more negative tuples lowering the rule confidence:

$$Age \in [18, 22] \rightarrow GraduatedCollege(40\%)$$

not to mention losing some of the rules focus. A tighter interval may boost confidence but lower support and vice versa.

Srikant and Agrawal attempt to counter this by mining across all values or intervals, which have been coded with a new integer representation, and then selectively combining them to increase support, producing additional, more general "items" as there are called. For example, $Price \in [500, 999]$ and $Price \in [1000, 1499]$ may both be probed for support in the database, leading to

$Price \in [500, 1499]$ being examined as well. All three could have minsupp and be used to find rules. The obvious problem here is that such a sweeping approach will incur much more computing time. They try to relieve this by adding a *maximum support* constraint to limit the combinations, though a lot of computing time will still be needed overall. The other less apparent caveat is the output of an excessive number of rules which could be either redundant or simply not useful. For example,

$$Price \in [500, 1499] \Rightarrow \quad Mem \in [64, 128] \quad (75\%sup, 80\%conf)$$
$$Price \in [500, 999] \Rightarrow \quad Mem \in [64, 128] \quad (25\%sup, 78\%conf)$$

shows that the second rule is essentially contained in the first and hardly conveys any more information. On this front Srikant and Agrawal have offered an automated *measure of interest* calculation to prune such redundancies before presentation to the user.

While quantitative rules do open up another kind of attribute to association rule mining, Kuok, Fu and Wong (1998) contend that there are still some limitations that need to be dealt with. Namely, the harsh boundaries created by the partitioning of quantitative attributes into intervals cause an unnatural division of the data and subsequently of the resultant rules. For example, the intervals $Age \in [20 - 29]$ and $Age \in [30 - 39]$ create an abrupt shift at the 29 year point, though in reality there is none. This is essentially the same problem discussed earlier regarding the term *RICH* and the crisp cutoff point ($\geq \$80,000$) for membership in that set. Kuok et al. (1998) offer a similar remedy for rule mining.

Quite simply, they propose fitting quantitative rules with fuzzy set theory to create *Fuzzy Association Rules*. Instead of the above Age intervals, the attribute Age could be partitioned into linguistic fuzzy sets still backed by the intervals at the membership function level. For example, we could have $Age \in young\ adult$ and $Age \in adult$, whose boundaries are fuzzy. Quantities of 21 - 28 could still have maximal membership (1.0) in the *young adult* set with membership beginning to decrease at 20 and 29. Ages 30 and above could also be members of the set but to a lesser and lesser degree. Likewise, ages 29 and below would be members of the *adult* set to some degree.

Kuok, Fu and Wong (1998) offer this formalism for their Fuzzy Association Rule technique (cf. Agrawal, Imielinski and Swami above):

T is transaction database of n tuples t: $T = \{t_1, t_2, \ldots, t_n\}$

$\mathcal{I}$ is the set of all items or attributes (e.g., Age, Price, etc.) in the database; $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$

Each item I_k can have membership in its own fuzzy sets:

$F_{I_k} = \{f^1_{I_k}, f^2_{I_k}, \cdots, f^m_{I_k}\}$, e.g.,

$$F_{age} = \{young\ adult, adult, middle - aged, senior\}$$

$$F_{height} = \{short, average, tall\}$$

A Fuzzy Association Rule is

$X\ is\ A \Rightarrow Y\ is\ B\ s.t.\ X \subseteq \mathcal{I} \wedge Y \subseteq \mathcal{I} \wedge X \cap Y = \emptyset$

Also

$A = \{f_{x_1}, f_{x_2}, \ldots, f_{x_p}\} \mid x_k \in X \wedge f_{x_k} \in F_{x_k}$

$B = \{f_{y_1}, f_{y_2}, \ldots, f_{y_q}\} \mid y_k \in Y \wedge f_{y_k} \in F_{y_k}$

Just as fuzzy set theory requires its own operations analogous to classical set operations, such as union, so too does fuzzy association rule mining. Boolean

association rules involve finding all *large* or *frequent itemsets* which meet a user-specified *support* threshold, support being the percentage of all database tuples which contain all elements of the itemset in question. Kuok et al. (1998) present an analogous measure called *significance* to gauge an itemsets frequency in the database. The major difference is, since we are dealing with fuzzy sets, each item's presence is weighted by its fuzzy membership degree in [0,1] before being divided by the number of tuples in the database T. In this respect, significance is similar to a fuzzy cardinality operation.

Formally, Kuok, Fu and Wong (1998) give the *significance factor*, $S_{\langle X,A\rangle}$ as:

$$Significance = \frac{\text{Sum of votes satisfying } \langle X, A\rangle}{\text{Number of tuples in T}}$$

$$S_{\langle X,A\rangle} = \frac{\sum_{t_i \in T} \prod_{x_j \in X} \{\alpha_{a_j}(t_i[x_j]))\}}{total(T)}$$

where

$$\alpha_{a_j}(t_i[x_j])) = \begin{cases} \mu_{a_j}(t_i[x_j]) & \text{iff } \mu_{a_j} \geq \omega \\ 0 & \text{otherwise} \end{cases} \quad a_j \in A$$

One thing to note here: the function α_{a_j} is a standard fuzzy operation known as an *alpha-cut* which is simply a threshold applied to a fuzzy set, in this case μ_{a_j}, that zeroes out memberships below a certain degree, in this case $0 < \omega \leq 1$.

Since we are using a product operation, any fuzzy membership that is zero, causes the measure for the entire tuple to be zero - the desired result. An unsupported attribute makes the entire attribute set useless for the significance calculation of that tuple. All tuple products, each having been weighted by the

attribute fuzzy values, are then summed and compared to the database as a whole, giving the fuzzy support or significance. Notice that if this operation were applied to crisp sets, as long as $0 < \omega \leq 1$ (which it should always be), the products will always end up as either 0 or 1 and the final result will be the classical itemset support as in the original work by Agrawal, Imielinski and Swami (1993).

If there is a fuzzy analogy to support, it stands to reason that there is also one for *confidence*. Kuok et al. (1998) offer a calculation for the *certainty factor* of a resultant rule, based on the above defined significance.

Not surprisingly, this factor is directly related to the confidence measures of classical association rules, which can be calculated on a rule from itemset Y:

$$conf_{X \Rightarrow Y-X} = \frac{supp(Y)}{supp(X)} \text{ where } X \subset Y$$

In like fashion, Kuok et al. (1998) offer:

$$\text{Certainty} = C_{\langle\langle X,A\rangle,\langle Y,B\rangle\rangle} = \frac{S_{\langle Z,C\rangle}}{S_{\langle X,A\rangle}} \text{ where } Z = X \cup Y, C = A \cup B$$

for a rule 'If X is A then Y is B', which is a logical extension of the support/significance analogy.

Through application of these techniques, Kuok, Fu and Wong believe they present a more robust system for mining useful real-world rules on quantitative data, utilizing fuzzy set theory operations.

2.4 Spatial Data Mining

According to Fayyad et al.(1996), *Data Mining* is the algorithmic application step of the full KDD (*Knowledge Discovery in Databases*) process which also

includes input preparation and output interpretation and utilization. Even so, it is a very broad term encompassing such subfields as clustering, neural networks, rule mining, etc. An important consideration when discussing data mining is the type of data being mined. General data mining literature usually centers on data typically found in a standard relational database system (RDBMS): retail transaction data, personnel data, client data, and the like. For the past several years, Jiawei Han and Kris Koperski have been among the chief exponents of *Spatial Data Mining* which they describe as

> ... the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases (Koperski and Han 1995).

A spatial database system is generally one like an RDBMS that has support for spatial data structures (e.g., MBRs - Minimal Bounding Rectangles), spatial operations (e.g., Spatial Join), and general spatial access methods or SAMs (Koperski, Han and Adhikary 1998). That said, a spatial database would not be exclusively tasked with handling spatial data, but in RDBMS fashion, would need to hold or have access to non-spatial data related to the spatial objects as well. Simply having a spatial object A and spatial object B is not as useful as knowing also that object A is a house and object B is a lake. In fact one of the uses for spatial data mining is "discovering relationships between spatial and nonspatial data" (Koperski, Han and Adhikary 1998).

Along these lines, Koperski and Han (1995) have done work on adapting the techniques of regular database association rule mining as presented by Agrawal, Imielinski and Swami (1993) for use with spatial data. Like the originals, spatial association rules still take the form $X \Rightarrow Y(c\%)$ except rather than

representing items or sets of items, X and Y now stand for spatial or nonspatial predicates(Koperski, Han and Adhikary 1998) such as $west\_of(x, y)$, spatial, since it deals with the objects location in space, or $isa(x, lake)$, nonspatial, since it describes the nature of the object but not its position. This is not far removed from the original association rule definition since they are *boolean rules* with *binary attributes* ("item X is present or it is not"). The aforementioned predicates are essentially the same thing. Rules involving various combinations of these could be mined to uncover hidden information about the data. For example,

$$\begin{aligned} isa(x, golfcourse) &\Rightarrow contains(x, lake)(100\%) \\ isa(x, town) &\Rightarrow adjacent\_to(x, river)(81\%) \\ close\_to(x, beach) \wedge isa(x, house) &\Rightarrow price(x, high)(76\%) \end{aligned}$$

are spatial association rules saying golf courses always have lakes inside, towns are usually located on rivers, and houses by the beach are usually expensive, "usually" being a linguistic approximation of the given high confidence values. It should be noted that the concept of support still applies, so while the golf course/lake connection may be strong, there may not be many golf courses in our database making the rule rather insignificant. A strong rule would have both high support and high confidence (Koperski, Han and Adhikary 1998). Koperski and Han use algorithms similar to those of Agrawal et al. to build the requisite large itemsets or predicatesets, integrating features required by the Spatial DBMS.

Another instance of spatial data mining of interest to us is the mining of image or raster data. Contrary to the earlier quote, spatial data mining does not have to be on an explicit spatial database but can be performed on a large image

database containing spatial data of a different kind, often from a remote sensing source. Koperski, Han and Adhikary (1998) say that this differs from ordinary "image processing" in scale:

> ... data mining studies very large amounts of data ... while image processing usually concentrates on analysis of single or a few images.

This kind of approach is applied when some kind of automated image analysis such as object detection or pattern matching is needed. A famous example is the work done by Burl et al. (1998) on recognizing volcanoes on the planet Venus. Mining on 30,000+ radar images from the Magellan probe, they were able to achieve an 80% accuracy rate with their system. The system consisted of three major components:

1. Data focusing to examine the image pixels (the building block of the images) and their neighbors to decide on the image areas to proceed on;
2. Feature extraction to glean necessary attributes and metrics (e.g., eigenvectors) from the found areas;
3. Classification learning which actually answers the question "volcano/not volcano" or learns from human expert examples how to do this based on the attributes from the last component.

This example is important because much of the spatial data coming in from field observation (e.g. satellite radar, aerial photos, seafloor SONAR, etc.) is indeed raw image data without the niceties of spatial objects structures, MBRs, etc. As such, much of the required work must be done here, at the image arrival point.

CHAPTER III

RESEARCH PLAN

3.1 Hypothesis

As was mentioned earlier, Bloch's fuzzy spatial techniques for 2D space can be applied to partitioned image data (i.e., containing discrete objects) to produce meta-data about the objects contained therein. This requires what is for the most part an implementation of Bloch's techniques, with the output customized for our purposes. After the meta-data has been extracted, we will apply association rule mining algorithms to these intermediate spatial relationship data to mine fuzzy association rules describing the *general* nature of the image components. However, as mentioned in the Literature Review, the existing fuzzy rule mining techniques (Kuok, Fu and Wong 1998) are designed for the general milieu of "traditional" non-spatial databases (e.g., age, price, etc.). Extensive modification of these will be required to properly mine our spatially derived meta-data to produce meaningful results. It is this adaptation and transformation of existing fuzzy rule techniques to mate with the fuzzy spatial relation detection methods in order to create a unified system to provide useful and interesting generalized information that is the focus of this inquiry.

Such a system will tackle some of the various unaddressed issues in the previously discussed literature via this integration. The fuzzy spatial relation work of Bloch usefully functions at the individual relation level, a specific object R related to a specific object A. The proposed system will mine these to extract

generalizations. The fuzzy association rules of Kuok, Fu and Wong add fuzziness to the classic associations of Agrawal, Imielinski and Swami, but principally apply it to ordinary non-spatial data. The proposed system will adapt these to a spatial/image domain. And while Koperski, Han and Adhikary do offer a spatial variation of association rule mining, it does not utilize fuzzy set theory. Indeed they cite a "(f)uzzy sets approach (1998)" as one of the many future directions of spatial data mining. The system proposed here takes up that challenge.

3.2 Methodology

We plan to use the fuzzy relative positioning techniques of Bloch (1999) to collect orientation meta-data about pairs of objects, previously detected by some other means. We can describe the scenario as:

Given:

$\mathcal{O}$ = $\{o \mid o$ is an object in $S\}$, set of all objects in space.

$\mathcal{C}$ = $\{c \mid c$ is the class of $o \in \mathcal{O}\}$, set of all object class labels.

The relation:

$$isa(o_x, c_i) \text{ where } o_x \in \mathcal{O} \wedge c_i \in \mathcal{C}$$

is a classical, crisp relation associating an object o_x with a class label c_i.

$$D_\alpha = \{((o_r, o_q), N_\alpha^{o_r}(o_q), M_\alpha^{o_r}(o_q), \prod_\alpha^{o_r}(o_q)) \mid (o_r, o_q) \in \mathcal{O} \times \mathcal{O} \wedge o_r \neq o_q\}$$

is a fuzzy relation, $D =$ "in direction of", or more specifically "o_q is in direction α of o_r." In contrast to typical fuzzy relation notation, we have explicitly replaced the single membership value, μ, with three values: the pessimistic necessity, N, the optimistic possibility, Π, and the overall mean, M. These are the same as those discussed on page 17 with $R = o_r$ and $A = o_q$ in this case. They have the property:

$$0.0 \leq N \leq M \leq \Pi \leq 1.0.$$

In other words, the necessity and possibility form an interval in which is found the Mean - $M \in [N, \Pi]$.

According to Bloch, this captures the inherent imprecision in the spatial relationship (1999).

For a given image-space, since we are examining the relationships between pairs of objects, we will find such tuples for all 2-permutations of the image's object set $\mathcal{O}$. Thus, the number of such tuples is:

$$P(|\mathcal{O}|, 2) = \frac{|\mathcal{O}|!}{(|\mathcal{O}| - 2)!} = |\mathcal{O}| \cdot (|\mathcal{O}| - 1)$$

Note that we want ordered permutations - not unordered combinations since the spatial relationships between objects are not commutative.

Furthermore, we will also not want to limit ourselves to a single α direction but rather several *primitive* directions, such as *right*, *above*, *left* and *below* or

$$A = \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$$

which can better describe the space in general.

Moreover, we will want to analyze not just an image but a set of images. This gives us a transaction database of $(|\mathcal{O}_i| \cdot (|\mathcal{O}_i| - 1) \cdot |A|)$ tuples for each image i, where $\mathcal{O}_i$ is the set of objects for image i. This is the simple, naive case; it may be possible to construct an effective pruning algorithm that obviates the need for full computation.

Given these tuples we will then apply previously discussed rule mining algorithms to them to extract rules. The exact nature of this mining, how field attributes are interpreted, and what the resultant rules will look like present a few difficulties.

How they are interpreted: While it is true that we may use natural numbers or alphabetic characters to denote classes, these are categorical attributes and not ordinal attributes. Class 2 is no closer to class 1 than class 7. As such, we

can interpret class as a categorical attribute as described in Srikant and Agrawal (1996).

The directional attributes may present somewhat more of a challenge. Fortunately, the work of Kuok, Fu and Wong (1998) establishes a convenient framework for interpreting direction. Kuok, Fu and Wong describe a set of fuzzy sets

$$F_{i_k} = \{f_{i_k}^1, f_{i_k}^2, f_{i_k}^3, f_{i_k}^n\}$$

that are associated with the attribute i_k in the set of attributes I. In our case, the set A of primitive directions represents a partitioning of the attribute *direction* aka D_α into such fuzzy sets

$$F_{D_\alpha} = \{f_{D_\alpha}^1, f_{D_\alpha}^2, f_{D_\alpha}^3, f_{D_\alpha}^4\} = \{Right, Above, Left, Below\}$$

The direction field of ordered pair of objects will have some degree of membership in each of these four fuzzy sets. Such an arrangement is visualized in Figure 3.1.

The question then arises as to what this membership degree is. Recall that we obtain three fuzzy membership values per pair per direction, necessity, mean, and possibility, instead of a simple μ value. The most likely candidate for use in rule mining, at least initially, is the mean, M, which is most akin to a single membership degree, being an average of such values. However, since we do generate the other two values which form the membership range of the spatial relation, they do have meaning and thus could be useful for mining. These could be split off to

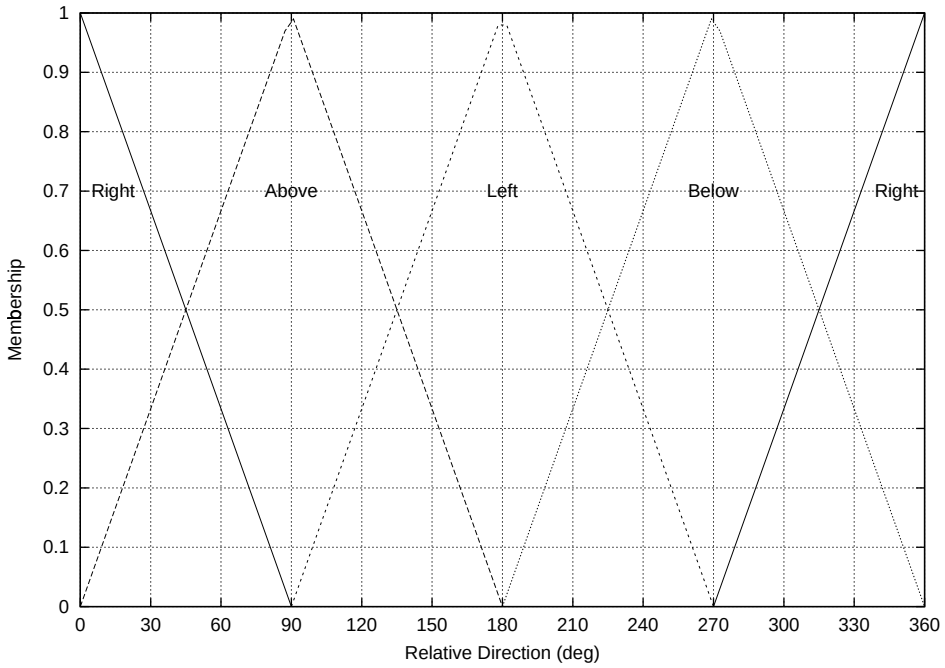

Figure 3.1: Primitive Direction Fuzzy Sets

form *sub-attributes* with their own analogous fuzzy sets:

$$
\begin{aligned}
F_{D_\alpha N} &= \{f^1_{D_\alpha N}, f^2_{D_\alpha N}, f^3_{D_\alpha N}, f^4_{D_\alpha N}\} \\
F_{D_\alpha M} &= \{f^1_{D_\alpha M}, f^2_{D_\alpha M}, f^3_{D_\alpha M}, f^4_{D_\alpha M}\} \\
F_{D_\alpha \Pi} &= \{f^1_{D_\alpha \Pi}, f^2_{D_\alpha \Pi}, f^3_{D_\alpha \Pi}, f^4_{D_\alpha \Pi}\} \\
&= \{Right, Above, Left, Below\}
\end{aligned}
$$

each one using the membership value indicated. Rules could then be generated taking some or all of these factors into account.

For the sake of consistency, one could also view the $isa(object_x, class_i)$ relation this way with sets for the class categorical attribute:

$$F_{class} = \{f^1_{class}, f^2_{class}, \ldots, f^n_{class}\} \text{ where } n = |\mathcal{C}|, \text{ the number of classes.}$$

Of course, *isa()* is a classical relation so the above actually represents the crisp set specialization of a fuzzy set where membership is in $\{0, 1\}$ rather than in $[0, 1]$. Also there is the further restriction that in our current approach, for a given tuple, the class attribute can have full membership in one and only one of these sets, e.g. $Object_{12}$ cannot be in $Class_3$ AND $Class_6$.

What the resultant rules will look like: If we adopt the attribute representation scheme of Kuok, Fu and Wong it stands to reason that we could adopt their rule representation as well. Recall their rule form was:

$$X \ is \ A \Rightarrow Y \ is \ B(s, c)$$

where X and Y are disjoint subsets of the set of all attributes I, A and B contain the corresponding sets for those attributes, and s and c are significance and certainty. An alternative form would be:

$$\forall x(isa(x, A) \Rightarrow (\exists y \ isa(y, B) \wedge D_\alpha(x, y, \text{membership} \geq \text{threshold})(\text{s, c})$$

However, if we view $isa(x, y)$ in the way noted above, this is really just an expansion of the form of Kuok, Fu and Wong.

It is this form that we will use for the initial round of experiments. However, subsequent experiments may require an extension to this arrangement to handle multiple directions. The basic form of Kuok, Fu and Wong only allows for one fuzzy set per attribute. That is, given our attribute D_α in Y for example, the accompanying set B of fuzzy sets can contain only one of the fuzzy sets in $F_{D_\alpha} = \{Right, Above, Left, Below\}$. But we may want to find rules about multiple directions at the same time; one object may be both to the right of and

above another object at the same time. To work around this, instead of viewing direction as one attribute with four associated fuzzy sets, we could view it as four attributes D_{α_z} each with its own single fuzzy set in which it has some degree of membership. This would allow the attribute set Y to potentially contain several directional attributes simultaneously and give a rule taking them all into account. Such a rule might be:

$$\forall x(isa(x, A) \Rightarrow \quad (\exists y\ isa(y, B) \wedge \quad D_{\alpha_1}(x, y, \text{membership} \geq \text{threshold}) \wedge$$
$$D_{\alpha_2}(x, y, \text{membership} \geq \text{threshold})(\text{s, c})$$

where D_{α_1} is the "Right" direction attribute and D_{α_2} is the "Above" direction attribute.

A side effect of this extension is that if we do indeed separate the directions this way and also use the previously discussed membership partitioning for necessity and possibility, we could end up with direction becoming twelve separate, albeit related, attributes:

	Right	Above	Left	Below
Necessity	$D_{\alpha_0}N$	$D_{\alpha_{\pi/2}}N$	$D_{\alpha_\pi}N$	$D_{\alpha_{3\pi/2}}N$
Mean	$D_{\alpha_0}M$	$D_{\alpha_{\pi/2}}M$	$D_{\alpha_\pi}M$	$D_{\alpha_{3\pi/2}}M$
Possibility	$D_{\alpha_0}\Pi$	$D_{\alpha_{\pi/2}}\Pi$	$D_{\alpha_\pi}\Pi$	$D_{\alpha_{3\pi/2}}\Pi$

This expansion could make mining more difficult, but on the other hand it could allow for more expressive rules.

3.3 Plan of Attack

In order to test the hypothesis, an initial system will be constructed based on the methods discussed in the previous section. This system will be used as an experimental test-bed for development and refinement of those methods. The system will consist of a suite of modules to carry out the various tasks. Among these modules will be:

- Synthetic Image Data Generator (SIDG) - A program for generating images with known object distributions and relationships.
- Fuzzy Spatial Relation Detector (FSRD) - An implementation of Bloch's approach for fuzzy spatial relation detection in 2D space, chosen because it has been shown to produce good meta-data for our purposes. It will be applied to the images to produce tables of relation data, either as flat files or more likely relational database tables, based on the relations above. There is the potential problem of this method being computationally inefficient, but, along with her basic algorithm, Bloch offers a *propagation algorithm* which she claims is a good approximation of the regular one with a speed increase upwards of $20\times$. In actuality, this is not a major concern since the FSRD is just an intermediate step and not the main focus of this investigation.

  Code to actually detect and tag the objects in the raw image files will also be in this module.
- Fuzzy Association Rule Miner (FARM) - The heart of the system. This will use algorithms in the literature, such as those of Kuok, Fu and Wong (1998) or Srikant and Agrawal (1996) as a base for initial implementation. However, to properly operate on the tables of Bloch based spatial meta-data, extensive modification and redesign will be necessary to mine the desired spatial fuzzy association rules, as detailed below.
- Other general utilities, such as a relational database management system like Oracle or Postgres, as well as image file handling libraries will be incorporated as needed.

The results of the FSRD applied to the synthetic data will be rows of spatial relationships in a form like:

Object ID	class

and

Ref OID	Other OID	α	N	M	Π

With this preliminary system in place, an initial set of experiments will be run applying the system to increasingly complex and varying synthetic images. The SIDG will be used to create these synthetic images with known object distributions for analysis, starting with arrangements of simple shapes (e.g., squares). Subsequent sets will add more complex objects, e.g., L's, H's and/or O's of varying sizes. Next more organic and irregular objects will be used. The first program that will actually be run on this data will be the Bloch-based FSRD which will generate the tabular meta-data whose format was outlined above. Tables of this type will be the transaction database to be mined by the FARM module.

The results of this first trial will determine the path of subsequent iterations of the system. The challenges of results analysis include:

- how to calculate and then apply significance and certainty, aka support and confidence. The crisp measures of support and confidence are well documented in Agrawal, Imielinski and Swami (1993), Agrawal and Srikant (1994), and others. Kuok, Fu and Wong (1998) offer a modification as significance and certainty for their fuzzy rules. Are these sufficient for our purposes, or does the nature of mining on spatial relations rather than simple attributes require a more refined approach?
- how to weigh the actual membership values. Considering that relations can have membership in several fuzzy sets at the same time, might more attention have to be paid to the weights of these memberships overall. Should a membership normalization step, such as that proposed by Luo and Bridges (2000), be integrated into subsequent versions?
- rule correctness. None of the above matters so much if accurate rules are not being mined.
- rule usefulness and interestingness. Do the rules convey interesting - possibly surprising - information about the database? More generally speaking,

> how well can we define interestingness and usefulness? There are not necessarily any definitive quantitative metrics in this area. As is often the case in computer science, it depends on the domain, the nature of the data, and ultimately the judgement of the end-user. That is not to say that investigations into this aspect of rule mining have not been mounted. Matheus, Piatetsky-Shapiro and McNeil have offered some factors including "novelty, utility, relevance, and statistical significance (1996)" as well as "estimated benefit achievable through available actions (1996)." Some of these are quantitative and objective, such as statistical significance. Others are qualitative and subjective, the user being the arbiter of success.

Such analysis will indicate what revisions will have to be made in the system and the subsequent round of experiments. Given that the initial system will mine only on the mean, M, membership value (see page 17) and only on one of the four primitive directions at a time (see discussion in Section 3.2), the rules from the initial experiments will probably be rather limited. The previously discussed extensions may need to be implemented.

One other concern that should be mentioned is that of system efficiency. We will be performing several different steps in this investigation, each of which incurs its own costs. For example, Bloch offers a complexity estimate of $O(n_R n_A)$ where n_R = number of points in reference object R and n_A = number of points in other object A for her basic spatial relation computation. However, this is only for a single object pair in an image, whereas we will be working on multiple images per set, multiple objects per image, and multiple angles per pair ($|\mathcal{O}| \cdot (|\mathcal{O}| - 1) \cdot |A|$ pairs per image as mentioned earlier). That is just one step in the process. While Kuok, Fu and Wong do not give a specific big-O complexity estimate for their algorithm (1998), we do know that association rule mining in general is a non-trivial computation. That said, this is a preliminary investigation into fuzzy spatial association rules in images. As such, efficiency is not of prime concern beyond

experimental feasibility (i.e., can we run our tests in a reasonable amount of time). Rather rule expressibility is the focus, with the building of an efficient "real-world" production system left for subsequent research.

So the ultimate goal of the investigation will be to develop a fuzzy association rule miner/spatial relation synthesis system that will recover interesting and significant association rules describing known spatial patterns in the synthetic data. At that point a brief application of this final system to the real data derived from preprocessed NAVO sea-floor data can be done as a first foray into possible future work.

3.4 Publication Plan

The results of these experiments will be submitted for publication at one of the following conferences:

Conferences:

- International Conference on Knowledge Discovery and Data Mining
- IEEE International Fuzzy Systems Conference (FUZZ-IEEE)
- or one of the other national fuzzy logic conferences

Also submission for publication will be made to one of the following journals:

- *Fuzzy Sets and Systems* - ISSN: 0165-0114
- *IEEE Trans. on Fuzzy Systems* (T-FUZZ) - ISSN: 1063-6706
- *IEEE Trans. on Knowledge and Data Engineering* (T-KDE) - ISSN: 1041-4347
- *IEEE Trans. on Pattern Analysis and Machine Intelligence* (T-PAMI) - ISSN: 0162-8828
- *International Journal of Intelligent Systems* - ISSN: 0884-8173

REFERENCES

Agrawal, R., T. Imielinski, and A. Swami. 1993. Mining associations between sets of items in massive databases. In *Proceedings of the 1993 ACM SIGMOD Int'l Conference on Management of Data held in Washington, DC, May 26-28, 1993*, 207–216. ACM Press.

Agrawal, R., and R. Srikant. 1994. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases (VLDB'94) held in Santiago, Chile, September 12-15, 1994*, edited by J. B. Bocca, M. Jarke, and C. Zaniolo, 487–499. Morgan Kaufmann.

Bellman, R., and M. Giertz. 1973. On the analytic formalism of the theory of fuzzy sets. *Information Sciences* 5:149–156.

Black, M. 1937. Vagueness: An exercise in logical analysis. *Philosophy of Science* 4:427–455.

Bloch, I. 1999. Fuzzy relative position between objects in image processing: A morphological approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(7):657–664.

Brule, J. 1985. *Fuzzy systems - a tutorial.* http://life.csu.edu.au/complex/tutorials/fuzzy.html (Accessed 12 Jul 2000).

Burl, M. C., L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, and J. Aubele. 1998. Learning to recognize volcanoes on venus. *Machine Learning* 30:165–194.

Chen, M.-S., J. Han, and P. S. Yu. 1996. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8(6):866–883.

Cheung, D. W.-L., J. Han, V. Ng, and C. Y. Wong. 1996. Maintenance of discovered association rules in large databases: An incremental updating technique. In *Proceedings of the Twelfth International Conference on Data*

Engineering (ICDE'96) held in New Orleans, February 26 - March 1, 1996, edited by S. Y. W. Su, 106–114. IEEE Computer Society.

Dubois, D., and H. Prade. 1980. *Fuzzy sets and systems: Theory and applications*, Volume 144 of *Mathematics in Science and Engineering*. New York: Academic Press, Inc.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. 1996. *Advances in knowledge discovery and data mining*. Menlo Park, CA: AAAI/MIT Press.

Güting, R. H. 1994. An introduction to spatial database systems. *VLDB Journal* 3(4):357–399.

Han, J., K. Koperski, and N. Stefanovic. 1997. Geominer: A system prototype for spatial data mining. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data held in Tucson, Arizona, May 13-15, 1997*, edited by J. Peckham, 553–556. ACM Press.

Holsheimer, M., and A. Siebes. 1994. *Data mining: The search of knowledge in databases*. Amsterdam: CWI. Technical Report CS-R9406.

Knorr, E. M., and R. T. Ng. 1996. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Transactions on Knowledge and Data Engineering* 8(6):884–897.

Koperski, K., J. Adhikary, and J. Han. 1996. Knowledge discovery in spatial databases: Progress and challenges. In *Proceedings of the 1996 ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'96) held in Montréal, June 2, 1996*, 55–70. IRIS/Precarn.

Koperski, K., and J. Han. 1995. Discovery of spatial association rules in geographic information databases. In *Advances in Spatial Databases: Proceedings of the 4th International Symposium on Large Spatial Databases, SSD'95 held in Portland, Maine, August 6-9, 1995*, edited by M. J. Egenhofer and J. R. Herring, 47–66. Springer.

Koperski, K., J. Han, and J. Adhikary. 1998. Mining knowledge in geographical data. ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/geo_survey98.ps (Accessed 12 Jul 2000).

Koperski, K., J. Han, and N. Stefanovic. 1998. An efficient two-step method for classification of spatial data. In *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98) held in Vancouver, July 12-15, 1998*, 45–54.

Kuok, C. M., A. W.-C. Fu, and M. H. Wong. 1998. Mining fuzzy association rules in databases. *SIGMOD Record* 27(1):41–46.

Luo, J., and S. M. Bridges. 2000. Mining fuzzy association rules and frequency episodes for intrusion detection. *International Journal of Intelligent Systems* 15(8):687–703.

Matheus, C. J., G. Piatetsky-Shapiro, and D. McNeil. 1996. Selecting and reporting what is interesting. In *Advances in Knowledge Discovery and Data Mining*, edited by U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Chapter 20, 495–515. Menlo Park, CA: AAAI/MIT Press.

Miyajima, K., and A. Ralescu. 1994. Spatial organization in 2d segmented images: Representation and recognition of primitive spatial relations. *Fuzzy Sets and Systems* 65:225–236.

Ng, R. T., L. V. S. Lakshmanan, J. Han, and A. Pang. 1998. Exploratory mining and pruning optimizations of constrained associations rules. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data held in Seattle, Washington, June 2-4, 1998*, edited by L. M. Haas and A. Tiwary, 13–24. ACM Press.

Ramaswamy, S., S. Mahajan, and A. Silberschatz. 1998. On the discovery of interesting patterns in association rules. In *Proceedings of 24th International Conference on Very Large Data Bases (VLDB'98) held in New York, August 24-27, 1998*, edited by A. Gupta, O. Shmueli, and J. Widom, 368–379. Morgan Kaufmann.

Schmucker, K. J. 1984. *Fuzzy sets, natural language computations, and risk analysis*. Rockville, MD: Computer Science Press.

Silberschatz, A., and A. Tuzhilin. 1996. User-assisted knowledge discovery: How much should the user be involved. In *Proceedings of the 1996 ACM SIGMOD Workshop on Research Issues on Data Mining and*

Knowledge Discovery (DMKD'96) held in Montréal, June 2, 1996, 89–94. IRIS/Precarn.

Smith, G. B. 1999. User constraints in data mining. Mississippi State University Computer Science Dept.

Srikant, R., and R. Agrawal. 1995. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Databases (VLDB'95) held in Zurich, Switzerland, September 11-15, 1995*, edited by U. Dayal, P. M. D. Gray, and S. Nishio, 407–419. Morgan Kaufmann.

Srikant, R., and R. Agrawal. 1996. Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data held in Montréal, June 4-6, 1996*, edited by H. V. Jagadish and I. S. Mumick, 1–12. ACM Press.

Yen, J., and R. Langari. 1998. *Fuzzy logic: intelligence, control, and information.* Upper Saddle River, NJ: Prentice Hall.

Zadeh, L. A. 1965. Fuzzy sets. *Information and Control* 8(3):338–353.

Zadeh, L. A., K.-S. Fu, K. Tanaka, and M. Shimura, eds. 1975. *Fuzzy sets and their applications to cognitive and decision processes.* New York: Academic Press, Inc.

Zimmerman, H.-J. 1996. *Fuzzy set theory - and its applications*, 3rd ed. Boston: Kluwer Academic Publishers.