Database Mining in Spatial Databases

D.A.Bell, S.S. Anand and C.M. Shapcott
Department of Information Systems
Faculty of Informatics
University of Ulster at Jordanstown
Northern Ireland
email: {dbell, ssanand, shapcott.cm}@uk.ac.ulster.ujvax

Abstract

The past few years has seen an explosion in the amount of data stored in databases. Apart from textual data, there has been an increase in the amount of pictorial information being stored on computers. Within this data is a lot of implicit, potentially useful information that cannot be extracted manually, simply because of the size of the data and the scarcity of human resources. Thus, the need for automated discovery tools is clear and a lot of effort is being devoted to building such tools.

In this paper we present a method for discovery in spatial databases. A general framework for databases mining in relational databases based on evidential theory [ANAN93] is extended to incorporate spatial discovery. We discuss the role of the Dempster-Shafer orthogonal sum combination rule in such a framework.

Keywords : Database Mining, Spatial Reasoning, Evidential Theory, Dempster-Shafer Orthogonal Sum.

1. Introduction

The discovery of knowledge can come about in many ways. Various modes of discovery are claimed from various quarters - for example, by accident, by purposeful observation or by a combination of learning and application of learned rules.

One *systematic* method of accumulating knowledge is by "mining" the data held in large data collections and it is normally understood as being intentional but not pre-planned to any level of detail. Examples of *database mining* include the classification of potential customers of a travel agency into those that suit a particular holiday package or not, classification of potential customers by banks into potentially good or bad investments, disaster prediction from climatic databases e.g. hurricane databases [MAJO93] and characterisation of employees from a personnel database by finding associations between their different attributes. In the latter case the data mining process might generate the following rule:

$$(\text{Age} < 25) \wedge (\text{Income} > 10000) \rightarrow (\text{Car\_model} = \text{Sports}).$$

The output of the data mining process is thus in the form of useful and interesting *rules*.
However, although this is the usual meaning of the term Database Mining, it is not the only one. For example, knowledge discovery by Database Mining with a different pattern of usage is presented in [FAYY93a]. In this case the data explorers have a clear and more specific objective. The goal here is to "map the sky" - catalogue and analyse objects in sky images. It is not to simply "look for interesting things in the available astronomy database", where interestingness is defined in some application independent way. The rules being discovered are known to involve certain attributes, are discovered by using a training set, and then applied to all of the database.

In this application and others like it, e.g. the paleantological investigation in [BELL93], the subset of interesting knowledge is pre-defined. The search is quite finely focused so the goal in [FAYY93a] is to "identify, measure and catalogue the detected objects in the images into classes". This is clearly a more specific method than generalised rule generation from a database using all the attributes.

Now within the range of knowledge discovery activities between these two extremes there lie various other patterns of usage of the term. In this paper we look at an exploration mode which could be used by application such as astronomy as above. It uses methods of combining evidence of a spatial kind in order to generate knowledge from experimental data.

We discuss this class of knowledge discovery exploration within a framework which *was* developed with the goal of employing Evidential Reasoning [ANAN93] in the process of mining a database. By virtue of a theoretical result, the Dempster-Shafer orthogonal sum [GUAN91] can be applied to general Boolean algebras [BELL94].

The particular Boolean algebra of interest here is a spatial one due to March [MARC88]. It allows two (or more) dimensional shapes to be manipulated algebraically. This, in turn, permits the *direct* representation of objects of interest in images, and association of uncertainties with them (- in our case uncertainties related to the exact position of volcanoes on Venus). The combination of the evidence allowing a choice to be made between alternative sites of volcanoes is by means of the Dempster-Shafer orthogonal sum.

The point should not be overlooked, however, that this is just a representative application for reasoning about space in the presence of uncertainty. Similar applications can be found in many fields such as medicine, weather-forecasting, and even searching for pilots who have abandoned their crashing planes over remote and difficult terrain.

The rest of this paper is in the following format. The next section describes the example problem that motivated the work presented here. Section 3 briefly describes the Evidential Approach that may be employed to solve the problem. Section 4 discusses a general framework for database mining based on Evidential Theory used by the authors for database mining in relational databases while section 5 describes how this framework may be extended for use in database mining in spatial databases. Section 6 discusses the use of Evidence Theory in the combination of spatial evidence. The method described can clearly be adapted within the framework of section 5 for discovery of knowledge in spatial databases.

2. Motivating Application

We use an example application of knowledge discovery in spatial databases given by Fayyad et. al. [FAYY93]. Fayyad et al. present methods based on pattern recognition and machine learning to analyse images of Venus returned by the Magellan spacecraft. Magellan transmitted over 30,000 high resolution images of the surface of Venus. The amount of data transmitted exceeded the amount of data transmitted by all the earlier planetary probes put together. Clearly the size of the data made manual analysis impossible. The goal of the analysis was to locate and parameterise volcanoes on the surface of Venus.

The authors present a system, for identifying volcanoes, that has three basic components - Focus of Attention, Feature Extraction, Classification Learning. The idea of the first component is to increase the efficiency of the system by identifying areas of the image being analysed which have a likelihood of a volcano. Its function is to scan the image and pick out regions of the image where there is a possibility of finding a volcano. This is done by comparing the intensity of a central pixel with the estimated background mean intensity of its neighbourhood pixels. The remaining task is then to discriminate between volcanoes and false alarms caused by other objects on the surface of Venus causing intensity deviations. The second component extracts features from the data and the third component uses training examples in order to create a classifier which can discriminate between volcanoes and false alarms.

In the next section we discuss a method for tackling aspects of this example and similar spatial database mining examples within an evidential theory framework.

3. An Evidential Approach

Like most remote sensing applications, the Magellan-Venus probe transmitted images of the same object from different angles and at different resolutions. Each of these images can be thought of as a source of evidence about the existence of volcanoes on the Venus Surface. The evidence collected from an image can be combined with other pieces of evidence collected from different images to arrive at a single value of uncertainty associated with the existence of a volcano in particular area.

Similarly, when analysing a single image for focusing the attention of the algorithm to particular areas of the image, it is possible to use different grids to locate *potential foci* and then combine them to get an overall set of foci of attention.

Fayyad et al. point out that in multi-sensor data, the different data sets need to be mapped to reference the same point on the imaged surface. This, being an imprecise process is normally prone to errors known as *spatial errors*. Thus there is a need to have degrees of confidence associated with conclusions made using the images. Such information is normally provided by using probabilistic models such as Bayesian methods. Probabilistic models implicitly assume the independence of data being analysed but this is not a valid assumption in spatial data as more often than not, the pixel elements in the image are highly correlated.

Evidential Reasoning [GUAN91,GUAN92] is a generalisation of conventional probability in that it does not make an assumption of independence as a matter of course. In examples like the Magellan-Venus exploration, where such independence cannot be assumed, the Evidential Reasoning is a better model for representing and manipulating uncertainty than probabilistic models.

4. A General Evidential based Framework for Database Mining

Anand et. al. [ANAN93] have presented a general framework for database mining based on Evidential Theory. The framework is inherently parallel as it can be partitioned into parallel streams. This makes the algorithms within the framework parallelisable and therefore likely to be efficient for large data sets. For a particular database mining task a subset of all the operators in the framework constitute the discovery procedure. Thus, adding a new discovery procedure is equivalent to adding new operators to the framework.

The basic idea behind our framework is very simple. Given a relation R, different pieces of evidence the existence of a rule in the relation can be collected by considering different random samples of the database. These pieces of evidence need to be combined to get a value for the uncertainty, support and interestingness of the rule for the whole database.

We consider a Database Mining framework to consist of two main parts :

- A method for representing data and knowledge
- Methods for data manipulation for discovering knowledge

Figure 1 shows our general framework for Database Mining.

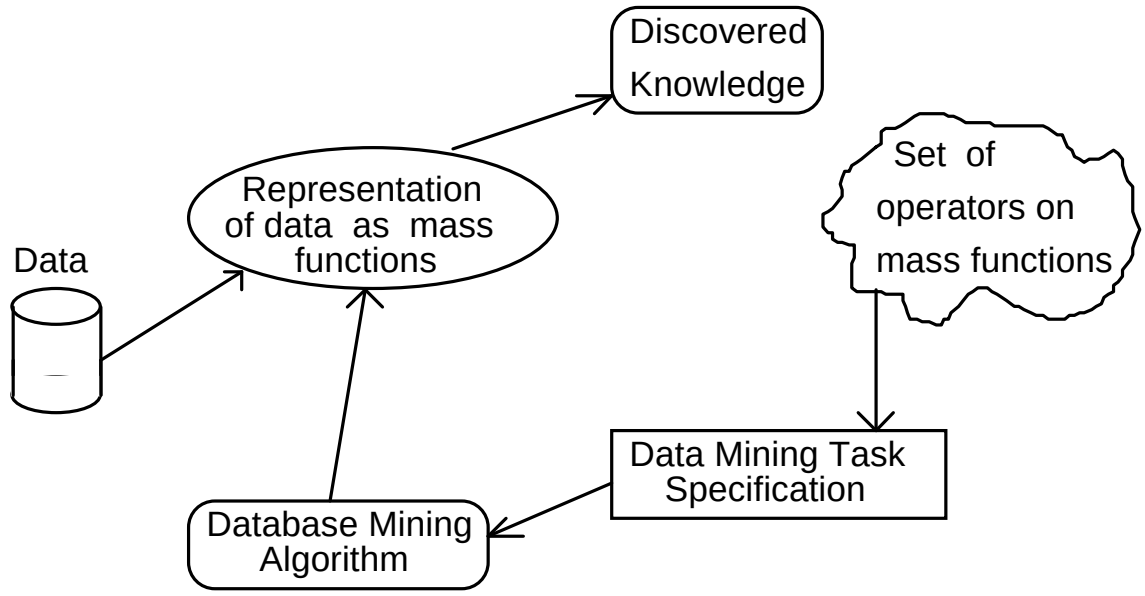

fig. 1 : General Framework for Database Mining

A rule induced from the database is of the form $A \rightarrow C$ where A is called the antecedent and C the consequent of the rule. Associated with each rule there are three measures :

- Uncertainty : The uncertainty associated with a rule is the ratio of the number of tuples in the database that satisfy both A and C to the number of tuples in the database that satisfy only the antecedent, A.
- Support : The support for a rule is the ratio of number of tuples of the database that satisfy both the antecedent and the consequent to the total number of tuples in the database.
- Interestingness : Clearly, the number of rules that can be extracted from large 'data mines' is probably as large if not larger than the actual amount of data in the database. We, therefore, need a method for measuring the degree of interest of a rule induced and only store a rule if its interestingness measure is greater than a threshold value. A number of indices for the interestingness of a rule have been suggested e.g. J-measure [SMYT91] based on Information Theory and Piatetsky-Shapiro [PIAT91].

We represent evidence pertaining to rules in the form of three *rule mass functions* defined as :

$$M : 2^A \text{ X } 2^C \rightarrow [0,1] \text{ X } [0,1] \text{ X } [0,1]$$

satisfying

1. $M(<\phi,\phi>) = (0,0,0)$
2. $\sum_{X \subseteq C,\, Y \in A} M[1](<Y,X>) = 1$

   $\sum_{Y \subseteq A,\, X \subseteq C} M[2](<Y,X>) = 1$

where A is the frame of discernment of Antecedents
and C is the frame of discernment of Consequents
and M[1],M[2] and M[3] are the first, second and third component of the 3-tuple associated with the rule by the mass function.
The framework considers the process of discovering knowledge to be a series of application of operators on rule mass functions. The framework classifies the operators into the following classes :

- **Combination operators** : These are binary operators which combine mass functions representing evidence of the existence of knowledge from different samples of the database to give one resultant mass function for the complete database e.g. the *Dempster-Shafer Orthogonal Sum.*

- **Statistical operators** : These are unary operators which are used on mass functions to deal with noise and missing values.
- **Induction operators** : These are unary operators which perform the process of inducing knowledge from the each of the database samples e.g. the *generalisation* operator in the STRIP algorithm [ANAN93a].
- **Domain operators** : These are unary operators which allow domain knowledge to be used in the discovery process e.g. the *coar* operator in the STRIP algorithm [ANAN93a].
- **Updating operators**

Given a particular database mining task, a set of operators can be selected to constitute the discovery procedure. Most of these listed above originate in database mining for relational databases, but this does not exclude applications in other non-textual systems. We now show how the above framework can be extended to discovery in spatial databases. In this case the database mining algorithm has quite a different flavour from that used in relational databases. Less emphasis is placed on probability calculation - estimates are assumed in this paper to be available from *some* source than on directed reasoning about spaces in the presence of uncertainty.

5. Extending the Framework to Spatial Databases

In a spatial database like the one consisting of data transmitted from the Magellan-Venus probe, data is stored in the form of images. Each image can be considered as a separate source of evidence of the location of a volcano on the Venus surface. Following the method used by Fayyad et. al. [FAYY93], the first task is to identify foci of attention within each image. We introduce the following technique based on Evidential Theory to achieve this.

5.1 Identifying Foci of Attention

The image being analysed is divided into a grid. Within each element of the grid, the intensity of the centre pixel is compared with the average background intensity of the rest of the element. If there is a high deviation in intensities, the grid element is taken to be a focus of interest. Using a number of different grids (a minimum of 2) we get different foci of interest. Combining these we get the foci of interest for the image. Associated with each focus of interest is a degree of confidence that a volcano exists in that area which can be calculated taking into account the intensity deviation within the focus.

5.2 Multi-Sensor Images

Images transmitted by the Magellan-Venus probe consisted of different images of the surface of Venus imaged with different resolutions and at different angles. Using the method described above, sets of foci of attention can be identified in each of these images - thus identifying potential volcanoes from the evidence provided by the image. These pieces of evidence can then be combined to get one overall set of foci of attention i.e. potential volcano sites.

Clearly, depending on the resolution of the image and the angle from which the image was captured, our faith in the evidence from the image will vary. In evidence theory, the *discounting operator* [GUAN93] performs this very function. Within our framework, the discounting operator would fall in the class of **Domain Operators** as it allows us to take into account biases on the strength of evidence available from particular sources (i.e. particular evidence). Without such an operator, the result of the combination would be erroneous.

The evidence provided by each image of the location of foci of attention can be combined using the Dempster-Shafer orthogonal sum operator to get an overall set of foci of attention with associated degrees of confidence.

6. Evidence Theory and Spatial Databases

6.1 Dempster-Shafer Evidential Reasoning

Evidence theory presupposes the existence of rules which allow the calculation of consequent attributes from antecedent attributes. It can be used in cases where the rules have already been generated using some form of machine learning as described above. For example geological experts might be asked to examine sampled images for volcanoes, providing a database of records consisting of features and volcanoes. Rules can then be extracted from this training database using the framework described above to allow the system to identify volcanoes from images in the rest of the database.

Assuming that rules have been extracted in this way, the Dempster-Shafer technique of evidence combination takes a set of pieces of evidence (the antecedents) as input and outputs a combined mass function which assigns a weight to each of the possible consequents. Thus in the volcano example, each image from a different sensor may be regarded as a piece of evidence which provides a certain amount of support for each of various possible locations of a volcano.

Evidence is represented in the Dempster-Shafer theory in the following way [GUAN91].

Let $< \chi, \cup, \cap, ', \phi, \Psi >$ be a Boolean algebra where

χ is the space of discernment,
$\cup, \cap$ and $'$ are the union, intersection and negation operations,
ϕ is the zero element and
Ψ is the greatest element in χ under set inclusion and χ is the power set of Ψ - the set of all subsets of Ψ.

A function m which maps elements of χ to the unit interval [0, 1] is called a *mass function* if :

1. It has non-zero values only at a finite number, F, A1, A2, ..., AF of members of χ,
2. $m(\phi) = 0$,
3. $\Sigma_{X \subseteq \Psi} m(X) = 1$

A function *bel* on χ is called a *belief function* if it can be expressed in terms of a mass function m:

$bel(A) = \Sigma_{X \subseteq A} m(X)$

To those who are accustomed to conventional probability theory a belief function, *bel*, corresponds to a l*ower bound* on the probability of a subset A of event space χ. Given a belief function *bel*, the function *dou*(A) = *bel*(A') is called a *doubt* function. The function *pls*(A) = 1 - *dou*(A) is called a *plausibility* function. Intuitively, the plausibility function provides an *upper bound* on the probability of subset A. Dempster-Shafer theory can be regarded as generalising conventional probability theory where probabilities are fixed and known to the case where only lower and upper bounds on probabilities are available.

When evidence is available from more than one source then each piece of evidence from a different source, i, is assigned its own mass function mi. Because a source may not be reliable, a given mass function may need to be *discounted*. After this discounting operation, the pieces of evidence can be combined using the Dempster-Shafer rule of evidence combination. The orthogonal sum operation $\oplus$ *combines* two mass functions m1 and m2 to yield a combined mass function $m1 \oplus m2$ as follows:

$$m1 \oplus m2\,(A) = \Sigma_{X \cap Y = A} \; m1(X) \cap m2(Y) \; / \; N$$

where the Normalising Constant N is defined by a double summation over all subsets of Ψ:

$$N = \Sigma_{X \cap Y \neq \phi} \; m1(X) \cap m2(Y)$$

The orthogonal sum operation can be applied repeatedly to cases where there are three or more sources of evidence in order to produce the combined weight of evidence.

In the situation where the frame of discernment Ψ is of a spatial nature then the Dempster-Shafer representation takes a form which is particularly easy to manipulate. In such cases the frame of discernment is represented by a grid and each element, A, of the space of discernment is simply a list of grid elements. This representation of spatial forms was developed by March [MARC88] in the context of built forms but can be generalised to any spatial problems where areas of interest can be represented by patterns on an underlying grid. As shown by March the list of G grid elements can be conveniently represented by a signature - an integer sig(A) consisting of G bits each of which is equal to one when the corresponding grid element is in the set A and zero if it isn't. As shown in previous work [BELL94] The Dempster-Shafer orthogonal sum as defined above requires the computation of set intersections such as $A1 \cap A2$ and these computations reduce to the bitwise AND of the corresponding signatures - sig(A1) AND sig(A2).

6.2 Example Using Spatial Evidence

Assume that there are several images, each image relating to the same underlying spatial grid. The problem is to determine the most likely location of a volcano on the grid. Each image is of the same area of the planet's surface but will have been created differently from the other images. This is because the space probe is moving and each image will have been taken at a different angle. An image consists of a grid of pixels, each pixel having a certain measured intensity. On its own the image can be analysed for evidence of the location of a volcano. The question is: "How should the evidence of several images be combined?"

Assume that the first reading produces the image shown in figure 2, suggesting that the volcano is located somewhere in the shaded region. However, the reading is not completely reliable. A rule generated by a training algorithm suggests that this evidence should be discounted by 20%.

$m1(A1) = 0.8$
$m1(\Psi) = 0.2$

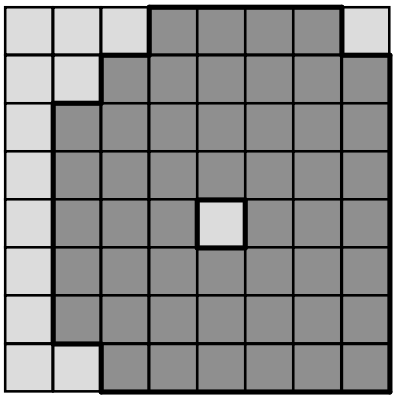

Fig. 2 : Image from Sensor 1

A1 is the set of shaded pixels which the sensor suggests may contain the volcano. Using a hexadecimal notation its signature is 003E 7FFF FFFF FF7F where pixels are numbered from top to starting at the top left of the diagram

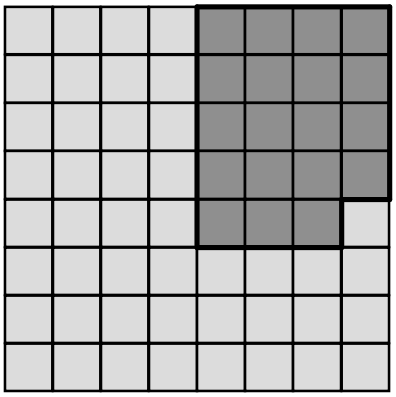

Fig. 3 : Image from Sensor 2

A second sensor has the associated diagram as shown in figure 3. The evidence of this sensor should be discounted by 35% , giving the mass function m2:

m2(A2) = 0.65
m2(Ψ) = 0.35
where A2 is the set of pixels which sensor two suggests contains the volcano.

Assuming that these two pieces of evidence can be combined their combined mass function can be found by the use of the Dempster-Shafer rule of combination in the following tabular form which lists all possible combinations of subsets with non-zero mass from the two mass functions:

m2 ⊕ m1	003E 7FFF FFFF FF7F (0.8)	FFFF FFFF FFFF FFFF (0.2)
0000 0000 F8F8 F8F0 (0.65)	0000 0000 F8F8 F870 (0.52)	0000 0000 F8F8 F8F0 (0.13)
FFFF FFFF FFFF FFFF (0.35)	003E 7FFF FFFF FF7F (0.28)	FFFF FFFF FFFF FFFF (0.07)

Suppose that a third sensor makes an additional mass function available which has the distribution outlined in figure 4. This sensor is more reliable than the other two and is discounted by 10%. This mass function has the form:
m3(A3) = 0.9
m3(Ψ) = 0.1

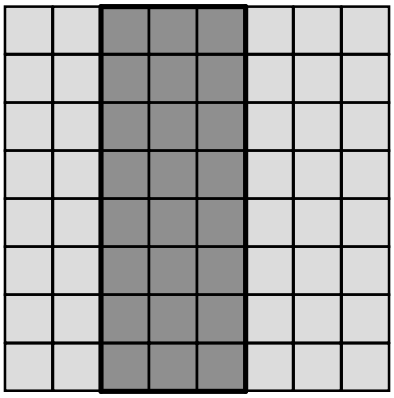

Fig. 4 : Image from Sensor 3

The orthogonal sum with the combined so far yields the combined evidence of the three sensors:

m12 ⊕ m3	0000 FFFF FF00 0000 (0.9)	FFFF FFFF FFFF FFFF (0.1)
0000 0000 F8F8 F870 (0.52)	0000 0000 F800 0000 (0.47)	0000 0000 F8F8 F870 (0.052)
003E 7FFF FFFF FF7F (0.28)	0000 7FFF FF00 0000 (0.25)	003E 7FFF FFFF FF7F (0.028)
0000 0000 F8F8 F8F0 (0.13)	0000 0000 F800 0000 (0.12)	0000 0000 F8F8 F8F0 (0.013)
FFFF FFFF FFFF FFFF (0.07)	0000 FFFF FF00 0000 (0.06)	FFFF FFFF FFFF FFFF (0.007)

The combined mass function thus derived concentrates the area in which the volcano is to be found.

This mass function can be summarised as follows:

```
m (0000 0000 F800 0000) = 0.47 + 0.12 = 0.585
m (0000 0000 F8F8 F870) = 0.052
m (0000 7FFF FF00 0000) = 0.25
m (0000 FFFF FF00 0000) = 0.06
m (0000 0000 F8F8 F8F0) = 0.013
m (FFFF FFFF FFFF FFFF) = 0.007
```

Hence, the three pieces of evidence from different sources have been combined to form a combined mass function which narrows the possible locations for the volcano.

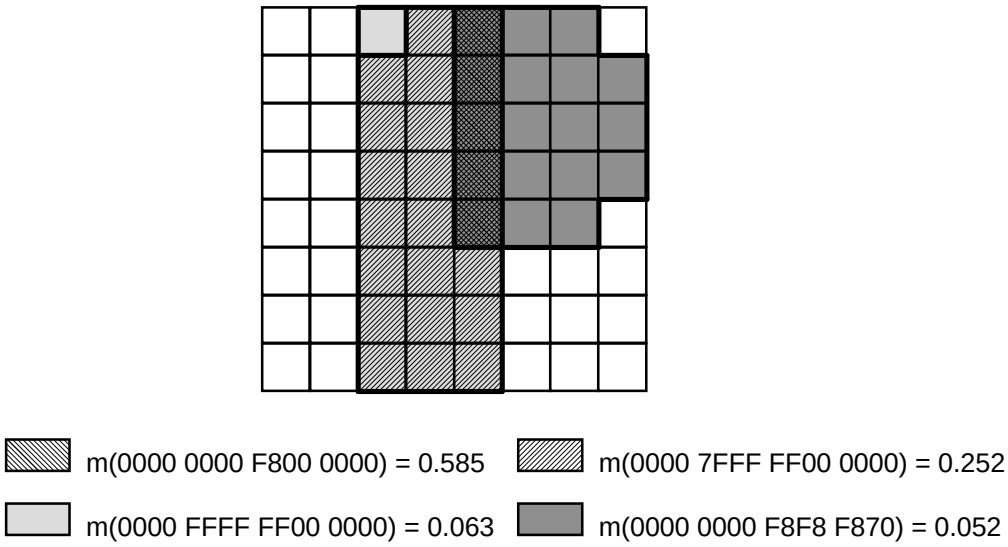

Fig. 5 : Graphical Representation of Combined Mass Function

Figure 5 shows the four largest components of the combined mass function, suggesting the most likely areas to contain the volcano.

7. Conclusion

We have presented a method for discovery in Spatial Databases based on Evidential Theory. The method presented here is an extension of a framework for discovery in relational databases. It uses the power of well established operators from Evidential Theory to combine spatial evidence from different sources.

The method employs the discounting operator and the well-known Dempster-Shafer orthogonal sum combination operator. Being a generalisation of Bayesian Statistics, Evidential Theory provides more honest results than conventional probability theory in that it does not make any implicit assumptions of independence.

Thus, this paper presents a strong basis for the creation of more powerful tools for discovery in spatial databases.

8. References

[ANAN93] Anand, S.S, Bell, D.A., Hughes, J.G. A General Framework for Database Mining Based on Evidential Theory Internal Report, Department of Information Systems, University of Ulster (Jordanstown), November, 1993.

[ANAN93a] S. S. Anand, D. A. Bell, J. G. Hughes Discovery of Strong Rules from Databases in Parallel Internal Report, Department of Information Systems, Univ. of Ulster at Jordanstown, Oct. 1993.

[BELL93] D. A. Bell From Data Properties to Evidence, IEEE Transactions on Knowledge and Data Engineering, Special Issue on Learning and Discovery in Knowledge - Based Databases, December, 1993.

[BELL94] D.A. Bell, J. Guan Using the Dempster-Shafer orthogonal Sum for Reasoning which Involves Space Internal Report, Department of Information Systems, University of Ulster at Jordanstown, March 1994.

[FAYY93] U.M. Fayyad, P. Smyth Image Database Exploration: Progress and Challenges Working Notes of the AAAI-93 Workshop on Knowledge Discovery in Databases, July 1993.

[FAYY93a] U.M. Fayyad, N. Weir, S. Djorgovski Automated Analysis of a Large-Scale Sky Survey: The SKICAT Survey Working Notes of the AAAI-93 Workshop on Knowledge Discovery in Databases, July 1993.

[GUAN91] J. Guan, D. Bell Evidence Theory and its Applications vol. 1 North-Holland, 1991.

[GUAN92] J. Guan, D. Bell Evidence Theory and its Applications vol. 2 North-Holland, 1992.

[GUAN93] J. Guan, D. Bell Discounting and Combination Operations in Evidential Reasoning Proc. of the Conference on Uncertainty in Artificial Intelligence, 1993.

[MARC88] L. March A Boolean Description of a Class of Built Forms in An Architecture of Forms ed. L. March, Cambridge Univ. Press, 1988.

[MAJO93] J.A. Major, J.J. Mangano Selecting Among Rules From A Hurricane Database Working Notes of the AAAI-93 Workshop on Knowledge Discovery in Databases, July 1993.

[PIAT91] G. Piatetsky-Shapiro Discovery, Analysis and Presentation of Strong Rules

Knowledge Discovery in Databases Pg. 229 - 248 AAAI/MIT Press 1991.

[SMYT91] P. Smyth, R. M. Goodman Rule Induction Using Information Theory Knowledge Discovery in Databases Pg. 159 - 176 AAAI/MIT Press 1991.