Open access • Journal Article • DOI:10.1007/S10707-010-0111-6

A framework for regional association rule mining and scoping in spatial datasets

Wei Ding, Christoph F. Eick, Xiaojing Yuan, Jing Wang ...+1 more authors

Institutions: University of Massachusetts Boston, University of Houston, University of Texas at Austin

Published on: 01 Jan 2011 - Geoinformatica (Springer US)

Topics: Association rule learning

Geoinformatica manuscript No.
(will be inserted by the editor)

A Framework for Regional Association Rule Mining and Scoping in Spatial Datasets

Wei Ding · Christoph F. Eick · Xiaojing Yuan · Jing Wang · Jean-Philippe Nicot

Received: date / Accepted: date

Abstract Advances in database and data acquisition technologies have resulted in an immense amount of spatial data, much of which cannot be readily explored using traditional data analysis techniques. The goal of spatial data mining is to automate the extraction of interesting and useful patterns that are not explicitly represented in spatial datasets. The motivation for regional association rule mining and scoping is driven by the facts that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional, rather than global. Furthermore, when using traditional association rule mining, regional patterns frequently fail to be discovered due to insufficient global confidence and/or support. This raises the challenges on how to measure the interestingness of a set of regions and how to search and evaluate regional patterns among those regions. This paper centers on

Preliminary versions of the paper appeared in [8, 7].

Wei Ding
Department of Computer Science
University of Massachusetts-Boston
Boston, MA 02125-3393 E-mail: ding@cs.umb.edu

Christoph F. Eick
Department of Computer Science
University of Houston
Houston, TX 77004
E-mail: ceick@uh.edu

Xiaojing Yuan
Engineering Technology Department
University of Houston
Houston, TX 77004
E-mail: xyuan@uh.edu

Jean-Philippe Nicot
Bureau of Economic Geology
John A. & Katherine G. Jackson School of Geosciences
The University of Texas at Austin
E-mail: jp.nicot@beg.utexas.edu

discovering regional association rules and determining the scope of these rules in spatial datasets. In particular, we present a reward-based region discovery method that employs clustering to find interesting places where regional association rules are valid. A divisive, grid-based supervised clustering algorithm is introduced for region discovery. We evaluate our approach in a real-world case study to identify spatial risk patterns from arsenic in the Texas water supply. Our experimental results not only confirm and validate research results in the study of arsenic contamination, but also lead to the discovery of novel findings that need to be further explored by domain scientists.

Keywords Region Discovery · Association Rule Mining and Scoping · Clustering · Spatial Data Mining

1 Introduction

Advances in database and data acquisition technologies have resulted in an immense amount of spatial data, much of which cannot be readily explored using traditional data analysis techniques. Techniques of spatial data mining have been developed to automatically find novel, useful, but implicit patterns from large spatial datasets [21, 36–38, 17, 12, 16, 26, 32, 34, 44]. Of particular interests to scientists is to find scientifically meaningful regions and their associated patterns, such as, identification of earthquake hot spots, association of particular cancers with environmental pollution, and detection of crime zones with unusual activities, etc.

The motivation for regional association rule mining and scoping is driven by the facts that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional, rather than global. It has been pointed out in the literature [15, 30, 35] that "*whole map statistics are seldom useful*," that "*most relationships in spatial data sets are geographically regional, rather than global*" and that "*there is no average place on the Earth's surface*" – a county is not a representative of a state, and a state is not a representative of a country. Therefore, it is not surprising that domain experts are most interested in discovering hidden patterns at a regional scale rather than a global scale [15, 28, 29].

Unfortunately, most of the current data mining techniques are ill-suited for discovering regional knowledge. For example, traditional association rule mining frequently fails to discover regional patterns due to insufficient global confidence and/or support. A common approach to alleviate the problem is to use a small support threshold. However, this approach usually suffers from a combinatorial explosion in the number of rules generated. Furthermore, for a given dataset, the number of regions as well as the regions themselves are not known *a priori*. This raises two questions: how to measure the interestingness of a set of regions and how to search for interesting regions. One popular approach is to select regions to be mined based on an *a priori* given structure, such as a grid structure using longitude and latitude, or based on political/demographical boundaries, such as counties within a state. However,

Well Depth	**Dangerous**	**Safe**	Total
(0, 251.5]	**1000**	1000	**2000**
(251.5, ∞)	1200	800	2000
Total	2200	1800	4000

	Well Depth	**Dangerous**	**Safe**	Total
ZoneA	(0, 251.5]	**400**	100	**500**
	(251.5, ∞)	1050	450	1500
ZoneB	(0, 251.5]	**600**	900	**1500**
	(251.5, ∞)	150	350	500
	Total	2200	1800	4000

Table 1 Contingency tables between well depth and arsenic concentration.

the boundaries of the so-constructed regions usually do not match the surface boundaries of the interesting patterns, making them difficult to be discovered. Using results from our real-world case study, let us consider an association rule that suggests a well X, up to 251.5-feet deep, is associated with dangerous arsenic concentrations:

$$depth(X, 0 - 251.5) \rightarrow arsenic\_level(X, dangerous)$$

Table 1 describes the well data of a county that includes Zone A and Zone B. Assuming the minimum confidence threshold is 70%, the pattern would not have enough confidence ($\frac{1000}{2000} = 50\% < 70\%$ threshold) to be identified globally in the county. However, the same rule holds in Zone A because its confidence, $\frac{400}{500} = 80\%$, is above the 70% threshold. Notice that this rule does not hold in Zone B, due to its low confidence ($\frac{600}{1500} = 40\%$). Hence a well up to 251.5-feet deep is *positively* associated with high arsenic contamination in zone A, but is *negatively* associated with dangerous arsenic concentration in the combined dataset. This reversal of an association in the global dataset is also known as spatial heterogeneity [38] or Simpson's Paradox in statistics[11].

Another interesting phenomenon is that regional association rules, by definition, only hold in a subspace but not in the global space; therefore, regional association rules may only be discovered in a particular subspace of the global space. This fact leads to novel challenges for regional association mining and scoping: (1) region discovery: how to determine regions from which useful association rules can be extracted; and (2) regional rule scoping: how to identify the scope of regional association rules. In this paper, we formalize both problems and propose our reward-based framework that utilizes the duality between regional patterns and regions where the patterns are supported: regions are used to discover regional association rules, and then regional association rules are used to determine regions in which those association rules are valid. Such regions provide a quantitative measure of how significant a regional association rule is in the global space.

Our Contributions. In this paper, we propose a novel efficient framework for regional association rule mining and scoping. We present a reward-based

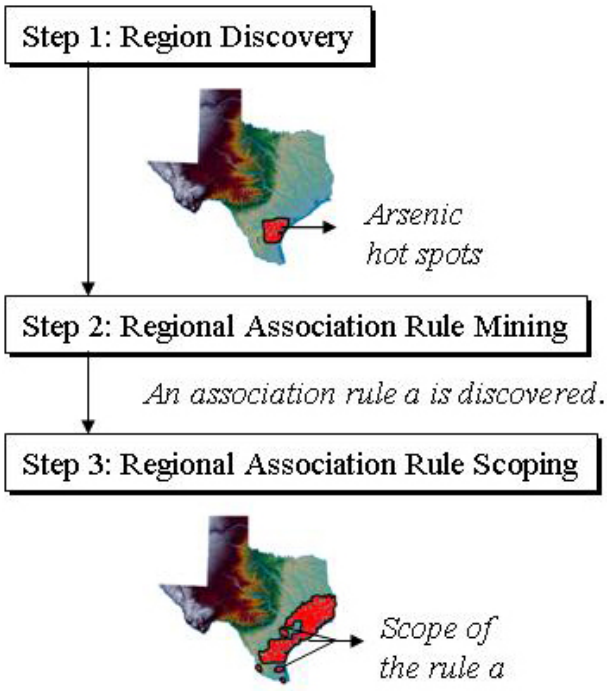

Fig. 1 An example for regional association rule mining and scoping.

region discovery method to search for regions that maximize an external fitness function, which captures what domain experts are interested in. Then an integrated approach is introduced to systematically mine regional rules from the discovered regions. At last we determine the scope of regional association rules to find places where those association rules are valid. We formulate region discovery as a clustering problem in which an externally given fitness function has to be maximized. Each cluster is assigned a "reward" value. A cluster receives a higher reward if a regional association rule exhibits stronger confidence and support.

We have designed and implemented a new divisive, grid-based supervised clustering algorithm to identify interesting regions in spatial datasets. The cluster algorithm searches for clusters to find interesting regions of arbitrary shape and scale.

We empirically evaluate the effectiveness of our framework using a real-world case study to identify spatial risk patterns from arsenic in the Texas water supply. Our experimental results not only confirm and validate research results in geoscience, but also lead to the discovery of novel findings that need to be further studied by domain scientists.

Figure 1 illustrates the basic procedure of our approach using a real example from our case study. An association rule a, *Wells with nitrate concen-*

trations lower than 0.085mg/l have dangerous arsenic concentration level, is discovered from an arsenic hot spot area in the South Texas with 100% confidence. The scope of the association rule a is a much larger area which mostly overlaps with the Texas Gulf Coast. Statistical analysis shows that the rule a cannot be discovered at the Texas state level due to its insufficient confidence (less than 50%).

The reminder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the framework of regional association rule mining and scoping. Section 4 describes the algorithms used in the framework. Section 5 presents the experimental results of a real-world application on identifying arsenic spatial risk patterns in the Texas water supply, and we conclude the paper in Section 6.

2 Related Work

The areas most relevant to our work are hot spot discovery, spatial association rule mining, spatial co-location pattern discovery.

2.1 Hot Spot Discovery

Hot spots are traditionally defined as clusters of "more than usual interest, activity, or popularity" with respect to spatial coordinates [25]. Wang *et al.* [44] introduce a "region-oriented" clustering algorithm to select hot spots to satisfy certain conditions such as density. Their approach uses statistical information, for example, means and standard deviations, instead of a fitness function to evaluate a cluster.

In spatial statistics, detection of hot spots using a variable resolution approach [4] was investigated in order to minimize the effects of spatial superposition. In [41], a region-growing method for hot spot discovery was described, which selects seed points first and then grows clusters from these seed points by adding neighbor points as long as a density threshold is satisfied. The definition of hot spots was extended in [22] using circular zones for multiple variables. Getis and Ord propose a popular method to find hot spots in spatial datasets relying on the G^* Statistic [14,31]. G^* Statistic detects local pockets of spatial association, and the value of G^* depends on an *a priori* given scale of the packets and is calculated for each object individually. Visualizing the results of G^* calculations graphically reveals hot spots (aggregates of objects with values of G^* higher than expected) and cold spots (aggregates of objects with values of G^* lower than expected). Note that such aggregates are not formally-defined clusters since the G^*-based method has no built-in clustering capabilities. Instead, hot spots are inferred from visualization and manual selection.

2.2 Spatial Association Rule Mining

Association rule mining has been introduced in [2] to mine interesting relations hidden in market basket transactions. Spatial association rule mining [21] extends association rule mining to spatial datasets. A spatial association rule takes the form of

$$P_1 \wedge P_2 \wedge ... \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge ... \wedge Q_n \ (sup\%,\ con\%).$$

It denotes an association relation among a set of predicates P_i $(i = 1, ..., m)$ and Q_j $(j = 1, ..., n)$, containing at least one spatial predicate. Spatial predicates may represent topological relations among spatial objects (e.g., intersecting, containing), or indicate a spatial orientation (e.g., north, left). The support of the rule ($sup\%$) measures the percentage of transactions containing both the antecedent and consequent of the rule. The confidence of the rule ($con\%$) indicates that $con\%$ of transactions that satisfy the antecedent of the rule will also satisfy the consequent of the rule. A rule $P_1 \wedge P_2 \wedge ... \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge ... \wedge Q_n$ is *strong* if $sup\%$ and $con\%$ satisfy minimum support and minimum confidence thresholds.

A common strategy used in spatial association rule mining is to divide the problem into three subtasks:

1. **Item representation and transaction definition**: define "items" and "transactions" for spatial datasets.
2. **Frequent itemset generation**: find all the itemsets that satisfy the minimum support threshold.
3. **Rule generation**: construct rules from the frequent itemsets that satisfy the minimum confidence threshold.

Apriori-style [2] association mining algorithms require that objects are described using categorical attributes. Therefore, continuous attributes have to be discretized which frequently results in information loss. Moreover, transactions are usually not *a priori* given for spatial datasets and therefore need to be defined. In addition, a transaction is not defined by nature in spatial space. If spatial association rule discovery is restricted to a reference feature (such as cities or wells), then transactions can be defined using the instances of this reference feature, as discussed in [21]. Our work adopts the same transaction model.

2.3 Spatial Co-Location Pattern Discovery

Shekhar *et al.* discussed several interesting approaches to mining co-location patterns, which are subsets of Boolean spatial features whose instances are frequently located together in close proximity [37,47,46]. Huang *et al.* proposed co-location mining involving rare events [17]. In [18], Huang and Zhang explored the relations between clustering and co-location mining. Instead of clustering spatial objects, the features of spatial objects are clustered using a

proximity function that is designed to find co-location patterns. However, it should be stressed that all the described approaches are restricted to global co-location patterns. Our approach, on the other hand, centers on discovering regional patterns. Aggarwal *et al.* [1] introduced localized association rule mining that seeks local association rules in clustered basket data. Their discovery is limited to non-spatial basket datasets, and they did not identify the scope of regional patterns.

3 The Framework for Regional Association Rule Mining and Scoping

As illustrated in Figure 1, the framework of regional association rule mining and scoping consists of three steps:

Step 1 Region Discovery: identifying interesting regions for regional association rules.
Step 2 Regional Association Rule Mining: mining regional association rules among discovered regions.
Step 3 Regional Association Rule Scoping: determining the scope of regional association rules.

In the remaining part of the section, we will first discuss our reward-based method for region discovery which is closely involved with Steps 1 and 3, and we will formally define the goal of our framework and formulate the measures of interestingness.

3.1 Region Discovery

Our region discovery method employs a reward-based evaluation scheme that evaluates the quality of the discovered regions. Given a set of regions $R = \{r_1, \ldots, r_k\}$ identified from a spatial dataset $O = \{o_1, \ldots, o_n\}$, the fitness of R, $q(R)$, is defined as the sum of the rewards obtained from each region r_j $(j = 1 \ldots k)$:

$$q(R) = \sum_{j=1}^{k} (i(r_j) \times size(r_j)^{\beta}) \tag{1}$$

where $i(r_j)$ is the interestingness measure of a region r_j, a quantity based on domain interest to reflect the degree to which the region is newsworthy. Our reward-based method seeks a set of regions R such that the sum of rewards over all of its constituent regions is maximized. $size(r_j)^{\beta}$ $(\beta > 1)$ in $q(R)$ increases the value of the fitness nonlinearly with respect to the number of objects in O belonging to the region r_j. A region reward is proportional to its interestingness, but given two regions with the same value of interestingness, a larger region receives a higher reward to reflect a preference given to larger regions.

We employ clustering algorithms for region discovery. A region is a contiguous subspace that contains a set of spatial objects such that for each pair of objects belonging to the same region, there always exists a path within this region that connects them. We search for regions $r_1, \ldots, r_k$ such that:

1. $r_i \cap r_j = \emptyset, i \neq j$, that is, the regions are disjoint.
2. $R = \{r_1, \ldots, r_k\}$ maximizes $q(R)$.
3. $r_1 \cup \ldots \cup r_k \subseteq O$: the generated regions are not required to be exhaustive with respect to the global dataset O.
4. $r_1, \ldots, r_k$ are ranked based on their reward values. Regions that receive no reward are discarded as outliers.

3.2 Problem Formulation

Let O be a spatial dataset, $S = \{s_1, s_2, ..., s_l\}$ be a set of spatial attributes, $A = \{a_1, a_2, ..., a_m\}$ a set of non-spatial attributes, and $CL = \{cl_1, cl_2, ..., cl_n\}$ a set of class labels. Let

$$\begin{aligned} I &= S \cup A \cup CL \\ &= \{s_1, s_2, ..., s_l, a_1, a_2, ..., a_m, cl_1, cl_2, ..., cl_n\} \end{aligned}$$

be the set of all the items in O, and let $T = \{t_1, t_2, ..., t_N\}$ be the set of all the transactions. T can be represented as a relational table, which contains N tuples conforming to the schema I (I contains $l+m+n$ items). An item $i \in I$ is a binary variable whose value is 1 if the item is presented in t_i ($i = 1, ..., N$) or 0, otherwise. Consequently, the set of transactions T is classified based on the given class structure CL.

Our framework leads to a class-guided generation of association rules that sheds more light on the patterns related to the given class structure. We define such rules as supervised association rules.

Definition 1 (Supervised Association Rule) A *supervised association rule* a is of the form $P \rightarrow Q$, where $P \subseteq I$, $Q \subseteq I$, *and* $(P \cup Q) \cap\ CL \neq \emptyset$.

The rule a holds in the O with confidence $conf$ and support sup where

$$sup(P \rightarrow Q) = \frac{|P \cup Q|}{N}$$
$$conf(P \rightarrow Q) = \frac{|P \cup Q|}{|P|}$$

where $|\ \ |$ denotes the number of elements in a set. A supervised association rule is *strong* if it satisfies user-specified minimum support (min_sup) and minimum confidence (min_conf) thresholds: $sup(P \rightarrow Q) \geq min\_sup$ and $conf(P \rightarrow Q) \geq min\_conf$.

The goal of regional association rule scoping is to compute a set of regions where a given association rule is valid. The scope of a regional association rule

represents the spatial impact of this regional pattern. We give formal definition of the scope of an association rule below.

Definition 2 (Scope of an Association Rule) The *scope* of an association rule a is a set of regions in which the association rule a satisfies the min_sup and min_conf thresholds.

Given these definition and nomenclature, the problem of regional association rule mining and scoping can be formulated as:

Find: interesting regions, supervised association rules from the discovered regions, and scope of strong regional association rules.
Given: a set of items I, a classified transaction set T, fitness functions for different measures of interestingness.

3.3 Measure of Interestingness

Different interestingness functions that correspond to various domain interests can easily be supported in our framework. In this section, we present two interestingness functions, $i_{hotpot\_coldspot}$ and i_{scope}, for regional association rule mining and scoping, respectively.

In function $i_{hotpot\_coldspot}$, the measure of interestingness is based on a set of class labels CL. It rewards regions whose probability distribution of CL significantly deviates from its priori probability. A region is a *hot spot*/*cold spot* if its probability distribution of CL is significantly higher / lower than an expected probability. The interestingness function $i_{hotpot\_coldspot}$ is calculated based on $P(r, CL)$ and $priori(CL)$, with the following parameters: η, γ_1, γ_2, R_+, R_-, where $\eta > 0$, $\gamma_1 \leq 1 \leq \gamma_2$, $0 \leq R_+, R_- \leq 1$. $P(r, CL)$ is the probability of objects in a region r belonging to CL, $priori(CL)$ is the probability of objects in datasets O belonging to CL, and R_+ and R_- are the maximum rewards for hot spots and cold spots, respectively.

$$i_{hotpot\_coldspot} = \begin{cases} \left[\frac{priori(CL)\times\gamma_1 - P(r,CL)}{priori(CL)\times\gamma_1} \times R_-\right]^{\eta} & if\ P(r,CL) < priori(CL)\times\gamma_1 \\ \left[\frac{P(r,CL) - priori(CL)\times\gamma_2}{1 - priori(CL)\times\gamma_2} \times R_+\right]^{\eta} & if\ P(r,CL) > priori(CL)\times\gamma_2 \\ 0 & otherwise \end{cases} \quad (2)$$

The parameter η determines how quickly the value of interestingness grows to the maximum value (either R_+ or R_-). If η is set to 1, the interestingness function changes linearly, as shown in Figure 2. In general, the larger the value for η is, the higher rewards for purer clusters are. $priori(CL) \times \gamma_1$ and $priori(CL) \times \gamma_2$ determine the thresholds based on which a reward is given to a cluster.

The following example explains how to calculate the fitness of a clustering schema X of an example dataset using Equations 1 and 2.

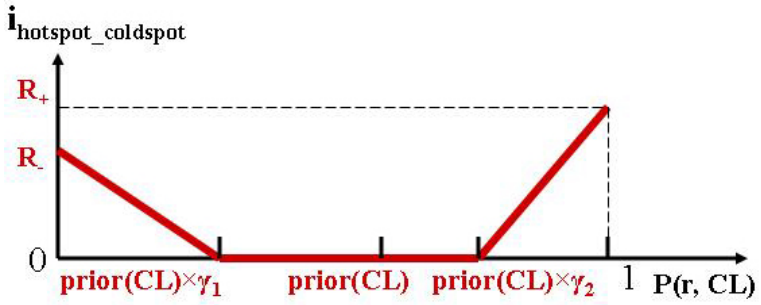

Fig. 2 The interestingness function $i_{hotpot\_coldspot}$ using $\eta = 1$.

Example Let us assume a clustering schema R is evaluated with respect to the class of interest *dangerous* (high-level arsenic) concentrations with $priori(dangerous) = 0.2$ and a dataset that contains 1000 examples. Suppose that the dataset is partitioned into 4 clusters $X = \{x_{11}, x_{12}, x_{13}, x_{14}\}$, and $|x_{11}| = 50$, $|x_{12}| = 200$, $|x_{13}| = 400$, $|x_{14}| = 350$. Assume that there are 20, 100, 80, and 0 objects labeled "dangerous" in the 4 clusters, respectively. $P(x_{11}, dangerous) = \frac{20}{50} = 0.4$, $P(x_{12}, dangerous) = \frac{100}{200} = 0.5$, $P(x_{13}, dangerous) = \frac{80}{400} = 0.2$, $P(x_{14}, dangerous) = \frac{0}{350} = 0$. The parameters used in the fitness function are as follows: $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 1$. Hence, $priori(CL) \times \gamma_1 = 0.2 \times 0.5 = 0.1$, and $priori(CL) \times \gamma_2 = 0.2 \times 1.5 = 0.3$. With this setting, a cluster does not receive any reward if its probability of class "dangerous" is not significanlty higher or lower than the expected probability, that is, the value is between $priori(CL) \times \gamma_1 = 0.1$ and $priori(CL) \times \gamma_2 = 0.3$. Therefore, x_{13} receives no reward. The interestingness for the other clusters using $\eta = 1$ are

$$i_{hotpot\_coldspot}(x_{11}) = (\frac{0.4 - 0.3}{1 - 0.3})^1 = \frac{1}{7},$$

$$i_{hotpot\_coldspot}(x_{12}) = (\frac{0.5 - 0.3}{1 - 0.3})^1 = \frac{2}{7},$$

$$i_{hotpot\_coldspot}(x_{14}) = (\frac{0.1 - 0}{0.1})^1 = 1.$$

The fitness value of the clustering schema X calculated using Equation 1 with $\beta = 1.1$ is

$$\begin{aligned} q(X) = \quad & \tfrac{1}{7} \times (\tfrac{50}{1000})^{1.1} + \tfrac{2}{7} \times (\tfrac{200}{1000})^{1.1} \quad + \\ & 0 \times (\tfrac{400}{1000})^{1.1} + 1 \times (\tfrac{350}{1000})^{1.1} \\ = \ & 0.369 \end{aligned}$$

Function i_{scope} evaluates the interestingness of a region for a given association rule. Let a be an association rule, $conf(a, r)$ the confidence of a in

a region r, and $sup(a, r)$ the support of a in r, we define the interestingness $i_{scope}(r)$ of a region r with respect to a given association rule a as follows:

$$i_{scope}(r) = \begin{cases} 0 & \text{if } sup(a,r) < min\_sup \times \delta_1 \text{ or} \\ & conf(a,r) < min\_conf \times \delta_2, \\ (\frac{sup(a,r)}{min\_sup})^{\eta_1} (\frac{conf(a,r) - min\_conf \times \delta_2}{1 - min\_conf \times \delta_2})^{\eta_2} & \text{otherwise.} \end{cases} \quad (3)$$

In regional association rule scoping, a region's reward is proportional to its interestingness, which is determined based on the confidence and support of association rule a in region r. In Equation 3, the thresholds $min\_sup \times \delta_1$ and $min\_conf \times \delta_2$ are introduced to weed out regions in which the association a barely holds. The minimum support and confidence thresholds prevent the clustering solution from containing large clusters with low interestingness. Values of parameters η_1 and η_2 ($\eta_1, \eta_2 > 0$) determine the weight to the increment of the support and confidence, respectively.

The measure of interestingness is designed to efficiently identify the scope of a given regional association rule. Firstly, in contrast to traditional association rule mining, the proposed measure of interestingness uses "soft" instead of "hard" thresholds to avoid a crisp effect [3]. For example, with $\delta_1 = \delta_2 = 0.9$, the function $i_{scope}(r)$ rewards regions as long as their confidence or support thresholds are within 90% of the hard thresholds min_conf and min_sup. Assume $min\_sup = 10\%$, $min\_conf = 80\%$, and an association rule whose support is 9% and confidence is 100% in a region r'. Instead of assigning zero reward to the region r', we argue to reward the region because the confidence of the rule is significantly above the min_conf threshold and its support is just a little bit lower (1%) than the min_sup threshold. Secondly, our approach uses a quantitative evaluation method that assigns a higher degree of interestingness and consequently a higher reward to regions whose support and confidence are high with respect to an association rule of interest. Thirdly, once an association rule a is discovered from a particular region r, we know that the region r from which the association rule a originates, receives a positive reward due to the fact that a satisfies the support and confidence thresholds in r.

4 Algorithms

4.1 Region Discovery

We formulate region discovery as a clustering problem to search for clusters that maximize domain-specific metrics as described in detail in previous section. Different measures of interestingness may lead to different sets of identified regions. Consequently, clustering algorithms embedded in the framework should allow for plug-in fitness functions. However, the use of fitness functions

is quite uncommon in clustering methods, although a few exceptions exist, for example, the hierarchical clustering algorithm CHAMELEON [20] uses fitness functions to evaluate inter-connectivity and proximity between two clusters. Furthermore, our region discovery method is different from traditional clustering methods as it is geared toward finding interesting places with respect to a given measure of interestingness. Clusters are ranked based on reward values, and clusters receive low reward are discarded as outlier and will not be identied as interesting regions.

We have designed and implemented a new Supervised Clustering algorithm using Multi-Resolution Grids (SCMRG). SCMRG is a hierarchical, grid-based method that utilizes a top-down search. The spatial space of the dataset is partitioned into grid cells. Each grid cell at a higher level is partitioned further into smaller cells at the lower level, and this process continues as long as the sum of the rewards of the lower level cells $q(R)$ is not decreased. The regions returned by SCMRG are the combination of grid cells obtained at different levels of resolution. The number of clusters, k, is calculated by the algorithm itself.

Algorithm 1 gives the pseudo-code of SCMRG. A queue data structure is used to store all the cells that need to be processed. The algorithm starts at a user-defined level of resolution and considers the following three cases when processing a cell c:

Case 1: if the cell c receives a reward, and its reward is greater than the sum of the rewards of its children ($succ(c)$) and greater than the sum of rewards of its grandchildren, this cell is returned as a cluster by the algorithm (steps 15-17).

Case 2: if the cell c does not receive a reward and its children and grandchildren do not receive a reward, neither the cell nor any of its descendents will be labeled clusters (steps 23-29).

Case 3: otherwise, put all the children of the cell c ($succ(c)$) into a queue for further processing (steps 18-21, steps 24-28).

The algorithm traverses through the hierarchical structure and examine those cells in the queue from the higher level. It uses a user-defined cell size as a depth boundary. Cells smaller than this cell size will not be split any further (step 19, step 25). Finally, SCMRG collects all the cells that have been identified in Case 1 from different levels, and merges neighbor clusters if it improves the fitness as defined in Equation 1. The obtained regions are returned as the result of executing SCMRG (steps 31-33).

This hierarchical grid-based approach captures clustering information associated with spatial cells without recourse to the individual objects because we do not drill down a cell if it does not look so promising (Case 2). SCMRG avoids time-consuming distance calculation because it uses the grid structure to define the neighborhood of objects. The computational complexity of SCMRG is thus linear in the number of grid cells processed, which is usually much less than the number of objects. Thus, the algorithm is capable of processing large datasets efficiently. The SCMRG algorithm has some similarity

Algorithm 1 The Algorithm of Supervised Clustering using Multi-Resolution Grids (SCMRG).

SCMRG (min_cell_size)
1. Determine a level of resolution l to start with.
2. Assign spatial objects to grid cells.
3. **for** each cell c at the current level l **do**
4. enqueue(c, $cellQueue$).
5. **end for**
6. **while** NOT $empty$(*cellQueue*) **do**
7. c = dequeue($cellQueue$).
8. r = reward (c). {Calculate reward for the cell.}
9. **for** each c_{child} ∈succ(c) **do**
10. $r_{children} = r_{children}$+reward ($c_{child}$).
11. **end for** {Calculate reward for its children.}
12. **for** each $c_{grandchild}$ ∈succ(succ(c)) **do**
13. $r_{grandchildren} = r_{grandchildren}$+reward ($c_{grandchild}$).
14. **end for** {Calculate reward for its grandchildren.}
15. **if** $r > 0$ {The cell receives a reward.}
16. **if** $r > r_{children}$ AND $r > r_{grandchildren}$
17. label the cell a cluster.
18. **else** {The cell should be divided further.}
19. **if** (the size of each c_{child} ∈succ(c) > min_cell_size)
20. enqueue($succ(c)$, $cellQueue$).
21. **end if**
22. **end if**
23. **else if** $r = 0$ {The cell does not receive a reward.)
24. **if** NOT ($r_{children} = 0$ AND $r_{grandchildren} = 0$)
25. **if** (the size of each c_{child} ∈succ(c) > min_cell_size)
26. enqueue($succ(c)$, $cellQueue$).
27. **end if**
28. **end if** {The cell should be divided further.}
29. **end if**
30. **end while**
31. Collect all the cluster-labeled cells from different levels.
32. Obtain regions by merging neighbor clusters if it improves the fitness.
33. Return the obtained regions.

with the STING clustering algorithm [44]. The difference is that the SCMRG algorithm focuses on finding interesting cells (those receive high rewards) instead of cells that contain answers to a given query. In addition, SCMRG only computes cell statistics when needed and not in advance as STING does, thus saving storage space as well.

The example in Figure 3 explains the procedure of this algorithm using a sample dataset. The first decomposition results into four cells $c_{11}, c_{12}, c_{13}, c_{14}$ at Level 1. If the reward of c_{11} is greater than the sum of the rewards of its children, and if it is also greater than the sum of rewards of its grandchildren, c_{11} is then labeled a cluster according to Case 1. Cell c_{14} does not receive any rewards, if neither its children nor grandchildren receive any rewards. According to Case 2, c_{14} is not labeled a cluster, and its successors are not saved in the queue. Although Cell c_{13} receives no reward, assume its children receive rewards, all the children of c_{13} are saved in the queue to be further processed (Case 3). The cells at Level 1 are then divided into Levels 2 and

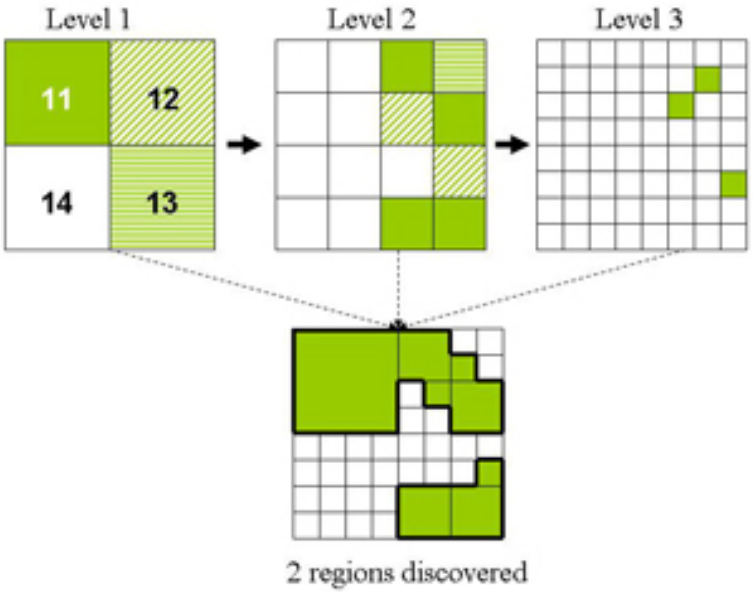

Fig. 3 Runing the SCMRG algorithm on a sample dataset.

3, and the same procedure is applied to all the cells in the queue. Each cell is labeled accordingly. The intermediate results are shown at Levels 2 and 3 in Figure 3. Neighbor clusters are merged if this improves the fitness. In this example, two regions are identified.

4.2 Generation of Regional Association Rules

Once regions are identified, we construct frequent itemsets for each region. Our Supervised_Apriori_Gen algorithm (see Algorithm 2) extends the *Apriori* algorithm [2] by utilizing a given class structure.

The *Apriori* algorithm first makes a single pass over the data set to determine the support of each single item, which generates all frequent 1-itemsets F_1. Next, the algorithm iteratively generates candidate k-itemsets using the frequent (k-1)-itemsets found in the previous iteration. A k-itemset is an itemset that has k attributes. A candidate itemset is pruned if it is not frequent. The algorithm terminates when there are no new frequent itemsets generated, for example, F_k=$\emptyset$. In Supervised_Apriori_Gen algorithm, the given class structure is incorporated by enforcing that each candidate k-itemset include at least one class label; otherwise it is pruned even if it is frequent. The Supervised-Apriori-Gen uses the $F_{k-1} \times F_{k-1}$ method [40] to merge a pair of frequent (k-1)-itemsets. Basically, let $A = \{a_1, a_2, ..., a_{k-1}\}$ and $B = \{b_1, b_2, ..., b_{k-1}\}$ be a pair of frequent (k-1)-itemsets. A and B are merged to form a k-itemset

Algorithm 2 Supervised_Apriori_Gen: Candidate Generation and Pruning.

```
Supervised_Apriori_Gen(F_{k-1})
1. if k = 2   {Deal with candidate 1- and 2-itemsets}
2.   for each frequent 1-itemset f ∈ F_1 do
3.     insert f into C_1.   {Generate candidate 1-itemsets}
4.   end for
5.   (C_{1_class_label}, C_{1_other}) = split(C_1, CL).
6.   for each candidate itemset c1 ∈ C_{1_class_label} do
7.     for each candidate itemset c2 ∈ C_{1_other} do
8.       c = form c1 and c2.
9.       insert c into C_2.   {Generate candidate 2-itemsets}
10.    end for
11.  end for
12.  for each candidate itemset c1 ∈ C_{1_class_label} do
13.    for each candidate itemset c2 ∈ C_{1_class_label} − {c1} do
14.      c = form c1 and c2.
15.      insert c into C_2.
16.    end for
17.  end for
18. else
19.  for each i1 in F_{k-1}
20.    for each i2 in F_{k-1}
21.      if (first k − 2 items of i1, i2 are same) ∧ (last item of i1, i2 differs)
22.        c = form (first k − 1 items of i1) and (last item of i2).
23.        insert c into C_k.
24.      end if
25.    end for
26.  end for
27. end if
28. return C_k.
```

$\{a_1, a_2, ..., a_{k-1}, b_{k-1}\}$ (see $form$ function in step 22) if they satisfy the following conditions:

$$a_i = b_i \qquad (for\ i = 1, 2, ..., k-2)\ \ and\ \ a_{k-1} \neq b_{k-1}.$$

The Supervised-Apriori-Gen algorithm initially starts with a candidate 2-itemset construction, which is the basis of the k-itemset generation ($k > 2$). To ensure that each 2-itemset includes at least one class label, the algorithm firstly constructs candidate 1-itemsets from frequent 1-itemsets (steps 2-4). The algorithm separates class-label items from other items using the *split* function (step 5). Next the algorithm enumerates class-label items with the rest of items (steps 6-11), as well as class-label items with themselves (steps 12-18). Thus, steps 6-11 generate candidate 2-itemsets formed between class labels and other non-class-label items; steps 12-17 generate candidate 2-itemsets formed between class labels. The 2-itemsets are then used for k-itemsets generation ($k > 2$) (steps 19-26).

After frequent itemsets are generated, we use the same approach proposed by the *Apriori* algorithm to generate strong supervised association rules using the min_conf threshold.

5 Arsenic Regional Association Rule Mining and Scoping in the Texas Water Supply

In this section, we describe the experimental procedures of applying the framework of regional association rule mining and scoping to a real world case study that identifies arsenic spatial risk patterns in the Texas water supply. We then discuss the experimental results and evaluate the performance of the proposed framework.

The experiments are conducted in four steps:

1. Data collection and data preprocessing, including cleaning data, transforming continuous attributes into categorical attributes, and constructing transactions using water wells as the reference feature.
2. Identifying arsenic *cold spots* and *hot spots*. A region whose arsenic distribution is significantly higher is considered an arsenic hot spot; a region whose arsenic distribution is significantly lower is considered an arsenic cold spot.
3. Mining supervised association rules from each identified region and for the complete dataset.
4. Determining scope of strong supervised association rules.

5.1 Datasets: Data Collection and Data Preprocessing

The datasets used in this study are extracted from the Texas Ground Water Database (GWDB) maintained by the Texas Water Development Board, the state agency in charge of statewide water planning [42]. The Texas Water Development Board has monitored and analyzed arsenic concentrations over the last 30 years. Arsenic in very high concentration is poisonous. Long term exposure to arsenic, even though at low level, can still lead to increased risk of cancers [39]. Arsenic is derived from both anthropogenic sources, such as the drainage from mines and mine tailings, pesticides, and biocides, and from natural sources, such as the hydrothermal leaching of arsenic-containing minerals or rocks. The World Health Organization has reported arsenic in drinking water in U.S., Thailand, Mexico, India, Hungary, Ghana, Chile, China, Bangladesh, and Argentina [45], as one of the key parameters for drinking water quality and safety evaluation.

Because data collection and maintenance procedures and standards have changed over the years in GWDB, datasets have to be cleaned to deal with problems such as missing values, inconsistent data, and duplicate entries. The obtained arsenic spatial dataset includes spatial attributes (S), non-spatial attributes (A), and class labels (CL) for each water well. Some of the spatial attributes are directly extracted from the database, such as *river basin*, *zone*, *latitude* and *longitude*. Implicit spatial attributes, such as *distance* between wells and rivers, are estimated using the 9-intersection model [10]. Non-spatial attributes are selected with the assistance of domain experts [19, 23, 33]; they include *well depth*, and concentration of *fluoride*, *nitrate*, and other chemical

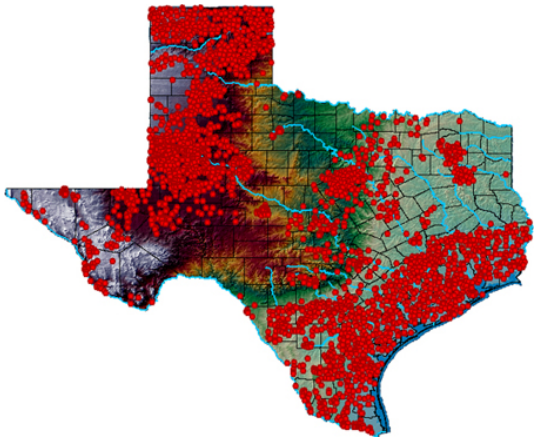

Fig. 4 Arsenic contamination in Texas; background depicts Texas terrain color ramp. Legend: red (or dark grey) dots – dangerous wells.

metal elements including *vanadium*, *iron*, *molybdenum* and *selenium*. Among those attributes, attribute *well depth* is used for studies on mobilizing mechanism; attributes *vanadium* and *molybdenum* have similar geochemical behavior; attributes *fluoride*, *nitrate*, *iron*, and *selenium* may suggest the ultimate origin of arsenic. The arsenic dataset generated by our research group is available on the web at [6].

We classify water wells into two classes: *safe* and *dangerous*. Based on the standard for drinking water defined by the Environment Protection Agency [43], a well is considered dangerous if its arsenic concentration level is above $10\mu g/l$. To ensure the quality of the association rule generated in the study, we only select lab test results that use honored sampling procedures. This results in 11,922 records selected from GWDB after data preprocessing. Figure 4 illustrates arsenic contamination in Texas, where dangerous wells are in red (or dark grey).

In preparation of the association rule mining, continuous attributes excluding latitude and longitude are first converted into categorical attributes. In general, two different methods are used for discretization of continuous attributes: unsupervised discretization without using class information and supervised discertization using class information [40]. In our experiments, we adopt the supervised method Recursive Minimal Entropy Partitioning introduced in [13]. The supervised entropy-based method uses class labels *dangerous* and *safe* to place the splits in a way that maximizes the purity of arsenic classes in the intervals. This discretization method maximized the support for arsenic class attribute, facilitating the discovery of supervised association rules involving with arsenic. Hence the method can effectively find the supervised association rules related with arsenic classes. The method produces unequal bin sizes and has been shown to produce better results in data mining tasks [9]. For example, the value of nitrate concentration has been discretized

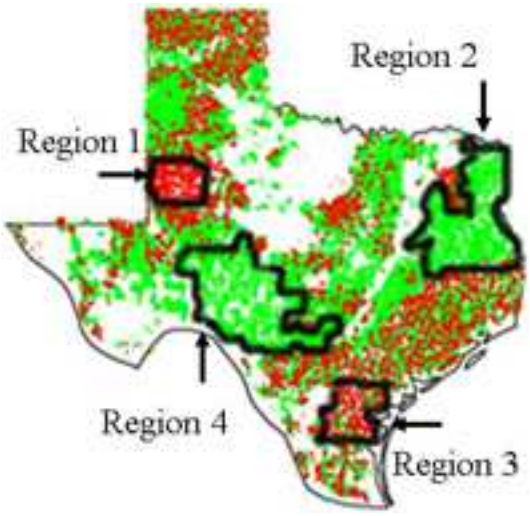

Fig. 5 Interesting regions are identified using $\beta = 1.01,\ \eta = 1,\ \gamma_1 = 0.5,\ \gamma_2 = 1.5,\ R+ = 1,\ R- = 1$. Average region purity = 0.85.

into five intervals with respect to the arsenic classes: (0,0.085], (0.085,0.455], (0.455,16.1], (16.1,28.085], and (28.085,∞) (measurement unit mg/l).

5.2 Experimental Results and Evaluation

We have re-discovered several interesting risk regions with high arsenic concentrations (hot spots), which have been studied by geoscientists before. We have also identified regions with low arsenic concentrations (cold spots). The association rules that we constructed from those identified regions can help geoscientists identify the causes of high arsenic concentrations in different regions. We now present our results with validation from the published results in geoscience for both region discovery and association rule mining and scoping.

In region discovery, the SCMRG algorithm is applied to a dataset that consists of longitude and latitude of wells along with arsenic class labels *dangerous* or *safe* using Equation 2. Figure 5 depicts the result of the top four regions that have received the highest reward. Specifically, Regions 1 and 3 have high density of dangerous wells, and Regions 2 and 4 have high density of safe wells. Hot spot Region 1 overlaps with the arsenic risk zone reported in the National Water-Quality Assessment Program [27], and hot spot Region 3 is confirmed as an arsenic risk zone by Parker's work [33].

If we are interested in finding larger regions with lower purity, using a larger value of β results in a bigger size of the regions. Figure 6 shows enlarged regions when β is increased from 1.01 to 1.035. In our experiments, we adjusted the granularity of regions by the quality of rules discovered in step 3. We observed

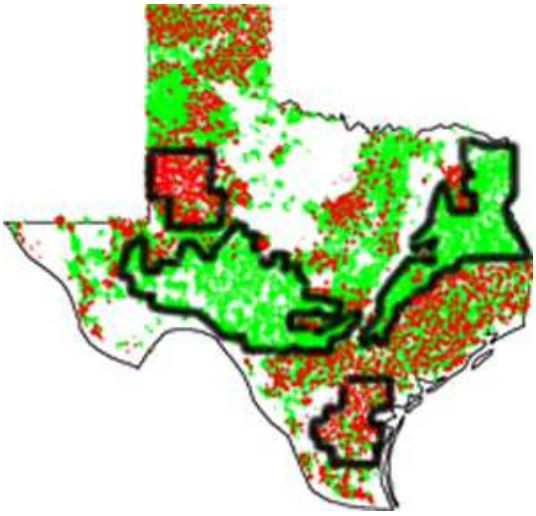

Fig. 6 Interesting regions are identified using $\beta = 1.035,\ \eta = 1,\ \gamma_1 = 0.5,\ \gamma_2 = 1.5,\ R+ = 1,\ R- = 1$. Average region purity = 0.83.

that $\beta = 1.01$ and $\eta = 1$ give us the best results in the rules constructed in supervised association rule mining.

The Supervised_Apriori_Gen algorithm is used to generate frequent itemsets for all the regions identified. We use $min\_support = 10\%$ and $min\_confidence = 70\%$ thresholds for the experiments. We present the first few rules for the regions investigated, which are all meaningful and important according to the arsenic study literature.

Mining regional rules in arsenic hot spots discovers attributes that are associated with high arsenic concentrations; in cold spots it discovers attributes related to low arsenic concentrations. For example, in Region 3 of Figure 5, we discover

$$is\_a(X, Well) \wedge nitrate(X, 0 - 0.085) \rightarrow arsenic\_level(X,\ dangerous)\ \ (100\%). \quad (1)$$

The rule states, with 100% confidence, that the wells in Region 3 with nitrate concentrations lower than $0.085 mg/l$ have dangerous arsenic concentrations. The strong association between nitrate and high arsenic concentrations is verified by Hudak's work [19] in environmental geology.

In Region 1 of Figure 5, we also discover

$$is\_a(X, Well) \wedge vanadium(X, 20.05 - 37.95) \wedge selenium(X, 74.55 - \infty) \rightarrow arsenic\_level(X,\ dangerous)\ \ (100\%). \quad (2)$$

The rule states with 100% confidence that the wells in Region 1, with vanadium concentrations between 20.05 and 37.95$\mu g/l$ and selenium concentrations larger than 74.55$\mu g/l$, have dangerous arsenic concentrations. Our discovery is confirmed by the work of Lee *et al.* in [23].

Our experiment results also show some novel rules that have not been reported in the literature of arsenic analysis. For example, in Region 1 the following rule is discovered:

$$is\_a(X, Well) \wedge depth(X, 0 - 215.5) \wedge iron(X, 19.65 - 20.05) \rightarrow arsenic\_level(X,\ dangerous)\ (100\%). \quad (3)$$

The rule indicates that shallow wells with a certain range of iron concentrations are associated with high arsenic concentrations. We hope that the results from our study will help domain experts in selecting interesting hypotheses for further scientific exploration.

Furthermore, we are interested to know whether the rules are different in different regions. We compared the sets of rules generated for Regions 1 and Region 3 (hot spots), as well as for Region 2 and Region 4 (cold spots). Due to different geographical structure and farm activities of the study area, the spatial risk patterns associated with arsenic are different in each region. For example, comparing the previously studied rule 1 identified in Region 3 with rule 4 extracted from Region 1:

$$is\_a(X, Well) \wedge nitrate(X, 28.085 - \infty) \wedge fluoride(X, 4.605 - \infty) \rightarrow arsenic\_level(X,\ dangerous)\ \ (100\%). \quad (4)$$

Instead of being related to relatively low concentrations of nitrate ($<$ 0.085mg/l), the rule says that with 100% confidence, the wells in Region 3, with high nitrate concentrations ($>$ 28.085mg/l) and fluoride concentrations higher than 4.605 mg/l, have dangerous arsenic concentrations.

Rules in Regions 2 and 4 (cold spots) shed light on what may prevent high arsenic concentrations. For example, we find the following rule, discovered both in Regions 2 and 4, states what is associated with low arsenic concentrations.

$$is\_a(X, Well) \wedge nitrate(X, 0.455 - 16.1) \wedge fluoride(X, 0.095 - 0.315) \wedge vanadium(X, 3.25 - 5.945) \rightarrow arsenic\_level(X,\ safe)\ (100\%) \quad (5)$$

In comparison, we also mine supervised association rules in the whole dataset. After some exploratory experiments, we found that by reducing the value of *min_support* from 10% to 1%, we are able to identify more interesting rules globally. However, in this case more than 100,000 rules are generated. It is painstaking to evaluate all these rules to find any meaningful ones. On the other hand, up to 300 rules on average are identified per region using our

regional rule mining. The need to use low support values for complete datasets has also been observed by [24]. However, all the regional rules (rules 1 to 5) that we discussed previously were not able to be identified due to low global confidence or support. Not surprisingly, statewide association rule mining finds very trivial and general rules, such as

$$is\_a(X, Well) \wedge water\_use(X, \text{“}by\ humam\ beings\text{”}) \wedge arsenic\_level(X, safe) \rightarrow inside(X, Basin19)\ \ (86\%)\ (6)$$

It claims that wells which are used by human beings and have safe arsenic concentrations are very likely (confidence is 86%) located in river basin 19 (in San Antonio area). It is a well-known fact in Texas.

We use the same clustering algorithm SCMRG but a different fitness function i_{scope}(Equation 3) for regional association rule scoping. The following four regional association rules with 100% confidence from Regions 1, 2, 3, and 4 are used as illustration examples in the rest of this section for regional association rule scoping. Association rules 1 and 3 are confirmed in arsenic literature [19, 23].

Association Rule 1

$$nitrate(X, 28.31 - \infty) \wedge arsenic\_level(X, dangerous) \rightarrow depth(X, 0 - 251.5)$$

Association Rule 2

$$depth(X, 0 - 251.5) \wedge fluoride(X, 0 - 0.085) \rightarrow arsenic\_level(X, safe)$$

Association Rule 3

$$nitrate(X, 0 - 0.085) \rightarrow arsenic\_level(X, dangerous)$$

Association Rule 4

$$depth(X, 251.5 - \infty) \wedge nitrate(X, 0.265 - 16.1) \rightarrow arsenic\_level(X, safe)$$

Figure 7 depicts the scope of four association rules above. The scope of an association rule can contain several regions. The scope of Association Rule 1 (top row, left column) overlaps with the Texas High Plains. In this area, shallow depth wells (< 251.5 feet) indicate the aquifer is thin; thus, nitrate comes from surface contamination ($> 28.31\ MG/L$). Arsenic contamination is of geological origin and is then enhanced by the lack of dilution because the aquifer is thin. The scope of Association Rule 3 (bottom row, left column) is applicable to the whole Texas Gulf Coast because the geology there is similar. The scope of Association Rules 2 and 4 represents the areas where arsenic contamination is low. They are interesting places that domain scientists will explore in the future.

It is also important to point out that the scope of an association rule indicates how global, regional, or local a pattern is. For example, the scope of

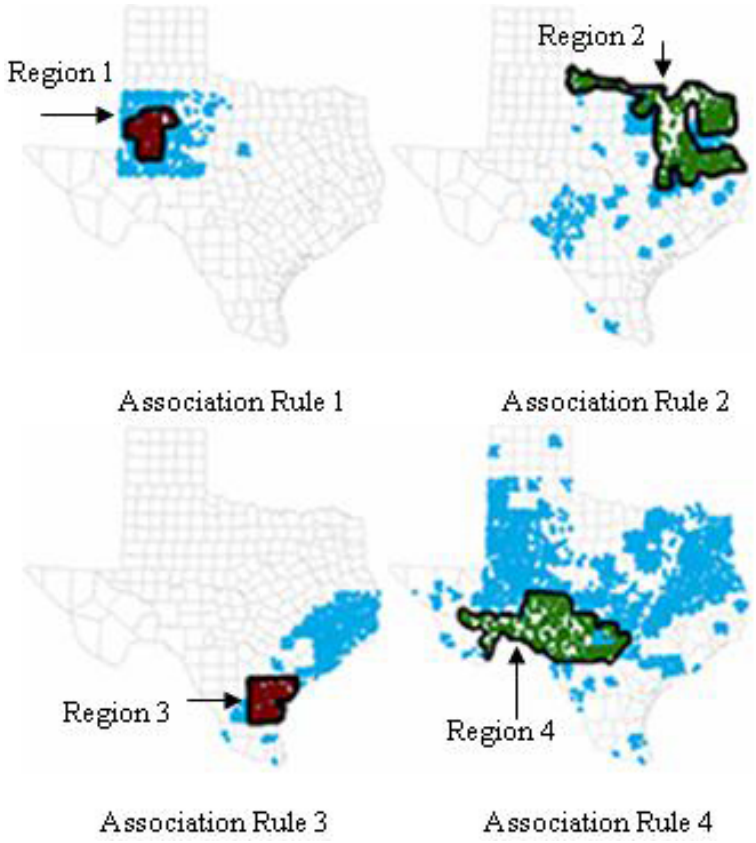

Fig. 7 Region - Regional association rule - Scope using $\beta = 1.01$, $\eta_1 = 1, \eta_2 = 1.1$, $\delta_1 = \delta_2 = 0.9$, $min\_sup = 10\%$, $min\_conf = 80\%$. Legend: regions are highlighted by bold border line; scopes are in color blue (or light grey).

the association rule 4 in Figure 7 covers a large percentage of the global space ($> 75\%$). We find that the association rule 4 is also valid (holds with 85% confidence) in the global dataset. Hence, it is indeed a global association rule. However, none of the other three association rules are discovered globally. We can also fine-tune the measure of interestingness for association rule scoping by varying its support and confidence thresholds for a given association rule. Figure 8 shows how the scope of the association rule 3 changes using different confience and support thresholds. Typically, a lower value of the min_sup results in a larger scope; a higher value of the min_conf results in a smaller scope.

Our SCMRG algorithm is computationally efficient. On average, it takes 3.031 seconds for hot spots/cold spots discovery, and 4.68 seconds for regional association rule scoping. The computer has an Intel(R) Pentium(R) M, a CPU 1.2GHz, and 632 MB of RAM. The algorithm can be accessed on the Web at

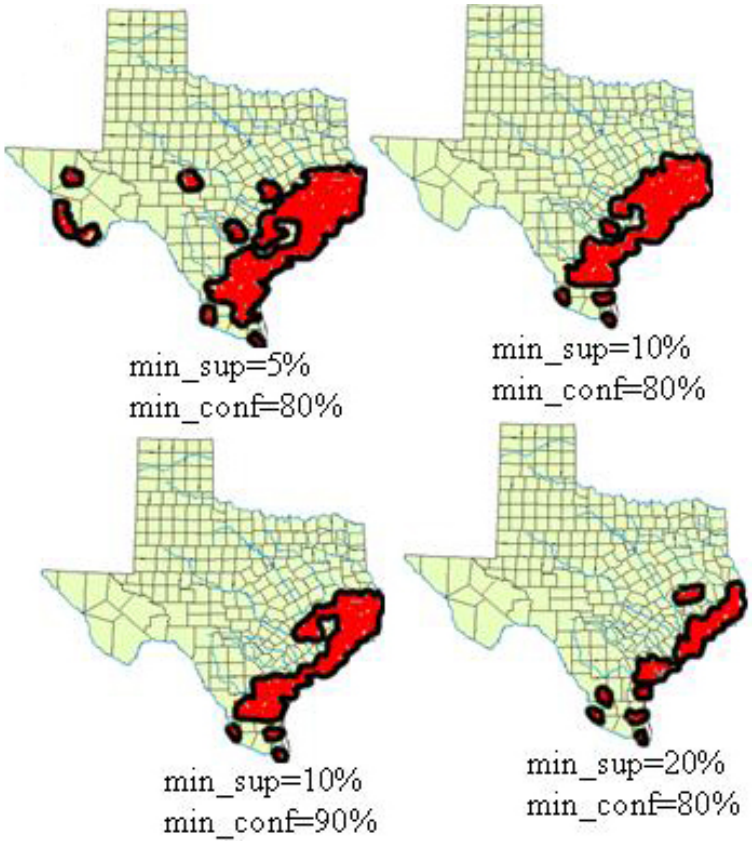

Fig. 8 The scope of a particular rule changes based on the different values of min_sup and min_conf. $\beta = 1.01,\ \eta_1 = 1, \eta_2 = 1.1,\ \delta_1 = \delta_2 = 0.9,\ min\_sup = 10\%,\ min\_conf = 80\%$.

our open source project *Cougar*[2] *Java Library for Machine Learning and Data Mining Algorithms* [5].

6 Discussion

One critical requirement for spatial data mining is the capability to analyze datasets at different levels of granularity, as well as analyze the data globally. We face two special challenges in regional association mining and scoping: (1) how to determine regions from which regional association rules will be extracted, and (2) how to evaluate the scope of regional association rules. We solve the first issue using a reward-based region discovery algorithm that employs a grid-based supervised approach to identify interesting subregions in spatial datasets. We address the second problem by exploiting the duality between regional patterns and regions: regions are used to discover regional association rules, then regional association rules are used to determine places

in which the association rules are valid. Such regions, defined as the scopes of regional patterns, provide a quantitative measure of how significant a regional association rule is in the global space. We evaluate the proposed framework in a real-world case study to identify spatial risk patterns and risk zones of arsenic in the Texas water supply. We have identified arsenic hot spots and cold spots, created regional rules from the obtained regions, and rediscovered associations that have already been reported in the scientific literature. Moreover, our approach identified several new relationships between arsenic and other factors which provide scientists with novel hypotheses for further exploration.

References

1. C. C. Aggarwal, C. M. Procopiuc, and P. S. Yu. Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering*, 14:51–62, 2002.
2. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.
3. S. Bistarelli and F. Bonchi. Interestingness is not a dichotomy: Introducing softness in constrained pattern mining. In *the Ninth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Lecture Notes in Computer Science*, volume 3721, Porto, Portugal, October 2005. Springer.
4. A. J. Brimicombe. Cluster detection in point event data having tendency towards spatially repetitive events. In *the 8th Intl. Conf. on GeoComputation*, 2005.
5. CougarSquared Data Mining and Machine Learning Framework, Data Mining and Machine Learning Group, University of Houston. *https://cougarsquared.dev.java.net/*, 2008.
6. Data Mining and Machine Learning Group, University of Houston. *http://www.tlc2.uh.edu/dmmlg/Datasets*, 2008.
7. W. Ding, C. F. Eick, J. Wang, and X. Yuan. A framework for regional association rule mining in spatial datasets. In *The 6th IEEE International Conference on Data Mining (ICDM)*, Dec. 2006.
8. W. Ding, C. F. Eick, X. Yuan, J. Wang, and J.-P. Nicot. On regional association rule scoping. In *the International workshop on Spatial and Spatio-temporal Data Mining in Cooperation with IEEE ICDM 2007*, Omaha, NE, USA, October 2007.
9. J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995.
10. M. J. Egenhofer and R. D. Franzosa. Pointset topological spatial relations. *International Journal for Geographical Information Systems*, 5(2):161–174, 1991.
11. S. EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, B13:238–241, 1951.
12. C. Eick, B. Vaezian, D. Jiang, and J. Wang. Discovering of interesting regions in spatial data sets using supervised cluster. In *PKDD'06, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006.
13. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In M. Kaufmann, editor, *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.
14. A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24:189–206, 1992.
15. M. F. Goodchild. The fundamental laws of GIScience. Invited talk at University Consortium for Geographic Information Science, University of California, Santa Barbara, 2003.
16. J. Han, M. Kamber, and A. K. H. Tung. Spatial clustering methods in data mining: A survey. In *Geographic Data Mining and Knowledge Discovery*, 2001.

17. Y. Huang, J. Pei, and H. Xiong. Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, 10(3):239–260, 2006.
18. Y. Huang and P. Zhang. On the relationships between clustering and spatial co-location pattern mining. In *ICTAI '06: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pages 513–522, 2006.
19. P. F. Hudak. Arsenic, nitrate, chloride and bromide contamination in the gulf coast aquifer, south-central Texas, USA. *Intl. Journal of Environmental Studies*, 60:123–133, 2003.
20. G. Karypis, E.-H. S. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.
21. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer and J. R. Herring, editors, *Proc. 4th Int. Symp. Advances in Spatial Databases, SSD*, volume 951, pages 47–66, 6–9 1995.
22. M. Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal Of The Royal Statistical Society Series A*, 164:61–72, 2001.
23. L. M. Lee and B. Herbert. A GIS survey of arsenic and other trace metals in groundwater resources of Texas. In *Natural Arsenic in Groundwater: Science, Regulation, and Health Implications (Posters)*, 2001.
24. W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *International Conference on Data Mining (ICDM'01)*, San Jose, CA, Nov. 2001.
25. Merriam-Webster Online Dictionary. *http://www.merriam-webster.com*, 2008.
26. R. Munro, S. Chawla, and P. Sun. Complex spatial relationships. In *The Third IEEE International Conference on Data Mining (ICDM)*, 2003.
27. National Water-Quality Assessment Program, U.S. Department of the Interior and U.S. Geological Survey. *Ground-Water Quality of the Southern High Plains Aquifer, Texas and New Mexico, Open-File Report 03-345*, 2001.
28. S. Openshaw. Two exploratory space-time attribute pattern analysers relevant to GIS. In S. Fotheringham and P. Rogerson, editors, *Spatial Analysis and GIS*, pages 83–104, London, 1994. Taylor and Francis.
29. S. Openshaw. Developing automated and smart spatial pattern exploration tools for geographical information systems applications. *The Statistician*, 44(1):3–16, 1995.
30. S. Openshaw. Geographical data mining: Key design issues. In *GeoComputation*, 1999.
31. J. K. Ord and A. Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286–306, 1995.
32. S. Papadimitriou, A. Gionis, P. Tsaparas, A. Visnen, H. Mannila, and C. Faloutsos. Parameter-free spatial data mining using MDL. In *5th International Conference on Data Mining (ICDM) 2005*, 2005.
33. R. Parker. Ground water discharge from mid-tertiary rhyolitic ash-rich sediments as the source of elevated arsenic in south texas surface waters. In *Natural Arsenic in Groundwater: Science, Regulation, and Health Implications*, 2001.
34. J. F. Roddick and M. Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. In *SIGKDD Explorations*, volume 1, pages 34–38, 1999.
35. S. Shekhar. Spatial data mining: Accomplishments and research needs. Keynote speech at GIScience 2004 (3rd Bi-annual International Conference on Geographic Information Science), 2004.
36. S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, (ISBN 013-017480-7), 2003.
37. S. Shekhar, Y. Huang, and H. Xiong. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions onKnowledge and Data Engineering*, 16:1472–1485, 2004.
38. S. Shekhar, P. Zhang, Y. Huang, and R. R. Vatsavai. Book chapter in data mining: Next generation challenges and future directions. In H. Kargupta and A. Joshi, editors, *Spatial Data Mining*. AAAI/MIT Press, 2003.
39. A. Smith and C. Hopenhayn-Rich. Cancer risks from arsenic in drinking water. In *Environmental Health Perspectives*, volume 97, pages 259–267, 1992.
40. P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

41. S. C. Tay, W. Hsu, and K. H. Lim. Spatial data mining: Clustering of hot spots and pattern recognition. In *IEEE International Geoscience and Remote Sensing Symposium*, 2003.
42. Texas Water Development Board. *http://www.twdb.state.tx.us/home/index.asp*, 2008.
43. U.S. Environmental Protection Agency. *http://www.epa.gov/*, 2008.
44. W. Wang, J. Yang, and R. R. Muntz. STING: A statistical information grid approach to spatial data mining. In *Twenty-Third International Conference on Very Large Data Bases*, pages 186–195, Athens, Greece, 1997. Morgan Kaufmann.
45. World Health Organization. *http://www.who.int/*, 2008.
46. H. Xiong, S. Shekhar, Y. Huang, V. Kumar, X. Ma, and J. S. Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *SIAM International Conf. on Data Mining (SDM)*, 2004.
47. J. S. Yoo and S. Shekhar. A join-less approach for mining spatial co-location patterns. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18, 2006.