Supervised Spatial Association Rule Mining: A Stratified Approach

Wei Ding
Computer Science Department
University of Houston
wding@uh.edu*

Christoph F. Eick
Computer Science Department
University of Houston
ceick@uh.edu †

ABSTRACT

In general, traditional association rule mining is unsupervised: transactions, from which association rules are constructed, do not belong to any particular classes. Supervised spatial association rule mining, on the other hand, aims to discover interesting, yet implicit rules from classified transactions. Furthermore, regional patterns in spatial dataset are often different from global ones, but frequently fail to be discovered due to insufficient global support. This problem is known as Simpson's Paradox. We propose a transaction stratification method using class labels to mine regional as well as global rules. Our framework leads to a class-focused generation of association rules that sheds more light on the patterns related to a given class structure. Our approach results in a more efficient discovery of relevant association rules and significantly reduces the number of rules generated. The proposed framework is evaluated with experiments on the Texas Ground Water Database to identify spatial patterns of risk from Arsenic.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications - Data Mining, Spatial databases and GIS

General Terms

Algorithms

Keywords

Spatial Association Rules, Stratification, Supervised Association Rules

1. INTRODUCTION

Association rule mining has been introduced in [2] to mine interesting relationships hidden in market basket transactions. Spatial association rule mining [15] extends association rule mining to spatial datasets. A spatial association rule takes the form of

$$X_1 \wedge X_2 \wedge ... \wedge X_m \rightarrow Y_1 \wedge Y_2 \wedge ... \wedge Y_n \quad (sup\%,\ con\%).$$

It denotes association relationships among a set of predicates X_i $(i = 1, ..., m)$ and Y_j $(j = 1..., n)$, where there exists at least one spatial predicate. Spatial predicates may represent topological relationships between spatial objects (e.g., intersects, contains), or indicate a spatial orientation (e.g., north, left). $sup\%$ is the support of the rule, which indicates that $sup\%$ of transactions contain both the antecedent and consequent of the rule. $con\%$ is the confidence of the rule, which indicates that $con\%$ of transactions that satisfy the antecedent of the rule will also satisfy the consequent of the rule. $X_1 \wedge X_2 \wedge ... \wedge X_m \rightarrow Y_1 \wedge Y_2 \wedge ... \wedge Y_n$ is *strong* if $sup\%$ and $con\%$ satisfy minimum support and minimum confidence thresholds.

In general, association rule mining is unsupervised: transactions, from which association rules are constructed, do not belong to particular classes. In this paper, we propose *supervised spatial association rule mining*, which aims to discover strong rules from classified transactions in spatial datasets. Our approach aims to mine class-focused association rules, which are defined as *supervised association rules*.

For example, we identified that 80% of wells, *inside river basin 14* and *less than 252 feet deep*, have *dangerous arsenic concentration level* from Texas Ground Water Database. 20% of all the wells satisfy these three above predicates, where *dangerous* is a class label and *inside* is a spatial predicate:

$$\begin{aligned} &is\_a(X, well) \wedge depth(X, 0-252) \wedge inside(X, Basin14) \\ &\quad \rightarrow arsenic\_level(X, class\_label : dangerous) \\ &\quad (20\%, 80\%). \end{aligned}$$

1.1 Spatial Association Rule Mining

A common strategy of spatial association rule mining is to decompose the problem into three subtasks:

1. Item representation and transaction definition: define "items" and "transactions" from spatial datasets.

2. Frequent itemset generation: find all the itemsets that satisfy the minimum support threshold.

*Also, Computer Science Department, University of Houston-Clear Lake, Texas
†$Revision: 2387 $

Table 1: A three-way contingency table between geographic zone A and zone B.

	Well Depth	Arsenic dangerous	Concentration safe	Total
ZoneA	(0,252]	40	10	50
	$(252,\infty)$	105	45	150
ZoneB	(0,252]	60	90	150
	$(252,\infty)$	15	35	50
		220	180	400

3. Rule generation: construct rules from the frequent itemsets that satisfy the minimum confidence threshold.

Association analysis requires item representation in binary form. Attributes can be continuous (e.g., longitude, latitude) or categorical (e.g., river basin). Continuous attributes need first be discretized into categorical attributes, which are then transformed into binary forms. Discretization may lead to following problems: when the intervals are too large, some rules may not satisfy minimum confidence; when they are too small, some rules may not satisfy minimum support [20]. Supervised spatial association rule mining utilizes class labels to set up bin boundaries to balance between the class purity and size of the intervals, see section 3.2.

Transaction definition is implicit 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 in [15]. Otherwise, transactions must be "invented" by mining algorithms, as in spatial co-location mining by [14] and spatial transactions by [19]. This paper adopts the transaction model in [15].

Classic algorithms such as Apriori [3] uses support-based pruning to systematically control the exponential growth of the candidate itemsets. Based on Apriori, our approach reduces the number of candidate itemsets by pruning the non-class relevant itemsets as early as possible, thus significantly reduce the computational complexity of frequent itemset generation. See discussion in section 3.4.

1.2 Stratification

One of the special characteristics of spatial data mining is: there is "no average place on the Earth's surface" [13]. Global patterns can be very different from local patterns. This phenomenon is known as Simpson's paradox [7]. We now explain the problem using table 1 and 2 to illustrate the relationship between well depth and arsenic concentration level for geographic zone A and B.

let's assume the minimum confidence is 70%. We now calculate the confidence of the following rule globally (zone A and Zone B) and locally (Zone A, Zone B respectively):

$$sample\_rule: \ is\_a(X, well) \wedge depth(X, 0-252) \\ \rightarrow arsenic\_level(danderous).$$

Table 2: A two-way contingency table between the well depth and arsenic concentration level.

Well Depth	Arsenic dangerous	Concentration safe	Total
(0,252]	100	100	200
$(252,\infty)$	120	80	200
	220	180	400

The rule states that a well up to 252 feet deep has dangerous arsenic concentration level.

Table 2 shows that the rule is not strong enough to be identified because its confidence 50% is less than the 70% minimum confidence:

$$\textbf{confidence}(sample\_rule) = 100/200 = 50\%.$$

In contrast, for zone A (Table 1) the rule is strong because its confidence 80% is above the 70% minimum confidence:

$$\textbf{confidence}(sample\_rule) = 40/50 = 80\%.$$

while this rule does not hold in zone B (Table 1):

$$\textbf{confidence}(sample\_rule) = 60/150 = 40\%.$$

The Simpson's paradox happens when there exist some underlying variables (in our example, zones) has a large effect on the ratios. Proper stratification proves to be an effective way to solve this problem [18]. Thus we propose a unique transaction stratification approach, see section 3.3.

1.3 Our Contributions

In this paper, we propose supervised spatial association rule mining, to mine association rules relying a given class structure. Furthermore, regional patterns in spatial dataset are often different from the global ones, but frequently fail to be discovered due to insufficient global support. We propose a transaction stratification method using class labels to mine regional as well as global rules. Our framework discovers class-focused association rules by utilizing class labels in data preprocessing, candidate itemset generation and pruning. This leads to a more efficient discovery of relevant association rules and significantly reduces the number of rules generated.

The outline of this paper is as follows: section 2 reviews related work; section 3 describes our approach; the experimental results are presented in section 4, and we conclude our study in section 5.

2. RELATED WORK

Previous works have proposed associative classification [16, 17, 22], which uses class association rules to build a more accurate classifier. Associative classification requires the consequent of a rule must be a class label. Our approach is different in that it uses classified transactions to guide the generation of association rules. In contrast to the classification association rule approach, our approach utilizes class labels for supervised attribute discretization, transaction stratification, and frequent itemset generation and pruning. Moreover, multiple class labels can be present either in the antecedent or consequent of a rule.

Many studies focus on the creation of "items" and "transactions" over spatial datasets so that Apriori-like [2] algorithms can be used. [15] uses two-step computation: first, association rules are generated at a coarse level, then only the spatial features with support higher than minimum support are passed to fine level rule generation. An interesting approach in [19] extracts spatial transactions based on the organization of GIS layers. On the other hand, [14, 23] extend the basket data transactions by defining spatial colocation patterns using event centric neighborhoods in place of transactions.

3. SUPERVISED SPATIAL ASSOCIATION RULE MINING

3.1 Basic Concepts

Let $\mathbb{D}$ be the spatial dataset, and $S = \{s_1, s_2, ..., s_l\}$ be the set of spatial attributes, $A = \{a_1, a_2, ..., a_m\}$ be the set of non-spatial attributes, and $CL = \{cl_1, cl_2, ..., cl_n\}$ be the 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 items in $\mathbb{D}$. Continuous attributes are transformed into categorical attributes, then all the categorical attributes are transformed into binary attributes. 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 following the schema I (I contains $l+m+n$ number of items). Thus an item $i \in I$ is a binary variable whose value is 1 if the item is present in t_i ($i = 1, ..., N$) and 0 otherwise.

The objective of supervised association rule mining is to discover strong relationships between class labels and other attributes. Formally,

Definition 1 A **supervised association rule** r is of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, *and* $(X \cup Y) \cap CL \neq \emptyset$.

The rule r holds in the $\mathbb{D}$ with confidence *con* and support *sup* where

$$\begin{aligned} sup(X \rightarrow Y) &= \frac{\sigma(X \cup Y)}{N}, \\ con(X \rightarrow Y) &= \frac{\sigma(X \cup Y)}{\sigma(X)}. \end{aligned}$$

The support count $\sigma(\alpha) = |\{t_i | \alpha \subseteq t_i,\ t_i \in T|$, where $|\ .\ |$ denotes the number of elements in a set. A supervised association rule is *strong* if it satisfies user-specified minimum support ($min\_support$) and minimum confidence ($min\_confidence$) thresholds.

We summarize the procedure of computing supervised association rule mining as follows:

1. Supervised discretization of continuous attributes. Continuous attributes are discretized using the class label set CL.
2. Transaction stratification. Transactions are stratified to find both regional and global association rules.
3. Frequent itemset generation. Candidate itemsets are generated and pruned using class label set CL.
4. Rule generation. Supervised association rules are extracted from the frequent itemset.

In the rest of section, we will talk about each step in details.

3.2 Supervised Discretization of Continuous Attributes

Because we are interested in mining supervised association rules, the best discretization approach should take class labels into consideration. We apply supervised discretization algorithm, Recursive Minimal Entropy Partitioning [11], to spatial and non-spatial continuous features. The algorithm uses entropy of candidate partitions along with class information to select boundaries for discretization. Discretization must avoid data fragmentation as too many small intervals may result in low support count. This is not a problem for entropy-based discretization since it is a global method [5]. The algorithm uses a top-down method to recursively partition a range of values into a set of intervals. The partitioning stops when $Gain(i, b; V)$ fails to satisfy a minimal entropy criterion, which is calculated by the number of class labels, number of instances, and entropy of each sets ($V,\ V_1,\ V_2$). Entropy-based information gain $Gain(i, b; V)$ is defined as:

$$Gain(i, b; V) = E(V) - (\frac{N_1}{N}E(V_1) + \frac{N_2}{N}E(V_2)),$$

where N is the number of instances in the value set V, V is all the possible values of item $i \in I$, b is a partition boundary, E is the entropy function. This approach measures the partitions along each branch of the recursive discretization independently. It is possible that some areas in the continuous spaces will be partitioned finely whereas others will be partitioned coarsely if it has low entropy.

3.3 Region-Based Transaction Stratification

In our work, we define a region R as a surface that contains a set of spatial objects. These spatial objects are the extension of R, denoted by $EXT(R)$. Moreover, we require regions be contiguous in spatial space – between any two objects in the same region, there exists at least one path that traverses only this region. However, in spatial data mining it is not only important to identify global patterns, but also to identify interesting regional patterns, because information is not uniformly distributed in spatial space [13]. Stratification is performed by stratifying transactions by identified interesting regions from a given spatial dataset. It has been pointed out in [12, 10] that proper stratification is necessary to avoid generating spurious patterns.

In this paper we propose the following supervised stratification approach. We assume a global space R and an underlying class structure C is given (objects in R are classified by C). Stratification starts with finding regions $R_1, ..., R_m$ such that:

1. $EXT(R_i) \subset EXT(R)$.
2. The regions are disjoint: $EXT(R_i) \cap EXT(R_j) = \emptyset$, $i \neq j$.
3. The spatial objects $EXT(R_i)$ in region R_i are pure or almost pure, that is, most of the objects in R_i belong to the same class, which is equivalent to $EXT(R_i)$ having a very low entropy with respect to the underlying class structure C.
4. The generated regions are not required to be exhaustive with respect to R, that is, $EXT(R_1) \cup ... \cup EXT(R_m) \subseteq EXT(R)$.

For example, for the wells in Texas depicted in Figure 2, our stratification approach identifies 4 sub-regions that are pure with respect to underlying class structure. Details will be given in the experimental evaluation.

In particular, we use a supervised clustering technique for region discovery, which generates a set of spatial clusters that correspond to regions. Transactions are then stratified by the set of regions, from which we mine regional rules. In previous work, we have developed an algorithm called Supervised Clustering using Multi-Resolution Grids (SCMRG)[8]. We adapt SCMRG in this paper to identify promising regions that will then be used to create regional association rules. Our approach employs a reward-based evaluation framework to measure the quality of different stratifications. The quality of a stratification of $EXT(R)$ into $X = \{EXT(R_1), ..., EXT(R_m)\}$ is computed as the sum of the rewards obtained from each $EXT(R_i)$ in X: the value of fitness function $fitness(X)$ is computed as the sum of the rewards obtained for each region $x \in X$.

$$fitness(X) = \sum_{x \in X} (reward(x)|c|^{\beta}),\ where\ \beta > 1.$$

Using $|c|^{\beta}$and $reward(x)$, $fitness(X)$ is optimized to find regions of the right size and opts to combining small regions into large ones, if the rewards of the combined regions are similar to that of the original two regions. SCMRG, itself, is a grid-based clustering algorithm that employs a divisive, top-down search. Each region at a higher level is partitioned further into smaller regions in the next level. The partitioning stops if the sum of the reward at the next level is not higher than the reward obtained at the previous level (See [8] for details). After regions are generated, transactions are stratified by the set of regions, and ready for frequent itemset generation.

3.4 Frequent Itemset Generation

In frequent itemset generation, we extend Apriori algorithm [2] on transactions by utilizing a give class structure. 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. Candidate itemset are pruned if it is not frequent. The algorithm terminates when there are no new frequent itemsets generated, it.e.g., F_k=Ø.

In our case, any candidate k-itemset must include at least one class label; otherwise it is pruned even if it is frequent. This is controlled by the candidate generation function called Supervised-Apriori-Gen (Figure 1). Supervised-Apriori-Gen uses the $F_{k-1} \times F_{k-1}$ method [21] to merge a pair of frequent (k-2)-itemset . 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)-itemset. A and B are merged if they satisfy the following conditions:

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

Supervised-Apriori-Gen initially starts with candidate 2-itemsets construction, which is the base of the K-itemsets generation ($k > 2$). First, the algorithm constructs candidate 1-itemsets from frequent 1-itemset (step 2-4). Second, to generate candidate 2-itemsets that focus on class labels, the algorithm separates class-label items from other items with *split* function (step 5). Then the algorithm enumerates class-label items with other items (step 6-11), as well as class-label items with themselves (step 12-17). Thus step 6-11 generate candidate 2-itemsets formed between class labels and other non-class-label items; step 12-17 generate candidate 2-itemsets formed between class labels. The 2-itemsets are then used for K-itemsets generation ($K > 2$) (step 19-20).

After frequent itemset generation, we use the same approach proposed by Apriori to generate strong supervised rules using $min\_confidence$.

4. EXPERIMENTS

We applied supervised spatial association rule mining to the Texas Ground Water Database (GWDB) to discover spatial patterns of risk from arsenic data in Texas. The purpose of the experiments is two-fold: finding the state-wide, as well as, regional spatial patterns of risk from arsenic. Arsenic, the 20th most popular element in nature, and is widely distributed throughout the earth's crust and commonly found in anthropogenic sources, such as drainage from mines, mine tailings, and pesticides. The population cancer risks due to arsenic in U.S. water supplies are comparable to those from environmental tobacco smoke and radon in homes [9].

GWDB is maintained by the Texas Water Development Board, the state agency in charge with statewide water planning [4]. Each well in the GWDB was treated as a transaction in our experiment. Raw attributes were transformed to create binary attributes. Because data collection methods and data maintenance have been changed over the years in the database, dataset has to be cleaned to fix missing values, inconsistent data, and duplicate entries in the dataset. Binary attributes in our experiments include spatial attributes, non-spatial attributes, and class labels. Some of the spatial features were directly extracted from the database, such as river basin, latitude and longitude. Implicit spatial features, such as distance between wells and rivers, were estimated using the 9-intersection model [6]. Non-spatial features are selected with the assistance of domain experts, such as well depth, zone, fluoride and nitrate concentration. We use ar-

Algorithm 1 Candidate Generation and Pruning: **Supervised_Apriori_Gen**

Supervised_Apriori_Gen(F_{k-1})
if $k = 2$ {Deal with candidate 1- and 2-itemsets}
  for each frequent 1-itemset $f \in F_1$ **do**
    insert f into C_1. {Generate candidate 1-itemsets}
  end for
  $(C_{1\_class\_label}, C_{1\_other}) = split(C_1, CL)$.
{Split C_1 , group class labels into $C_{1\_class\_label}$, and the other frequent 1-itemsets into C_{1_other}}.
  for each candidate itemset $c1 \in C_{1\_label}$ **do** {Generate candidate 2-itemsets with class-label items and non-class-label items}
    for each candidate itemset $c2 \in C_{1\_other}$ **do**
      $c =$ form $c1$ and $c2$.
      insert c into C_2. {Generate candidate 2-itemsets}
    end for
  end for
  for each candidate itemset $c1 \in C_{1\_label}$ **do**
    $C_{post} = subset\_split(C_{1\_label}, c1)$. {Identify all the class labels in the array C_{1_label} that is after $c1$}
    for each candidate itemset $c2 \in C_{post}$) **do**
      $c =$ form $c1$ and $c2$.
      insert c into C_2.
    end for
  end for
else
  for each $i1$ in count.
    for each $i2$ in F_{k-1}
      if (first $k-2$ items of $i1, i2$ same) $\wedge$(last item of $i1, i2$ differs)
        c =form (first $k-1$ items of $i1$) and (last item of $i2$).
        insert c into C_k
      end if
    end for
  end for
end if
return C_k

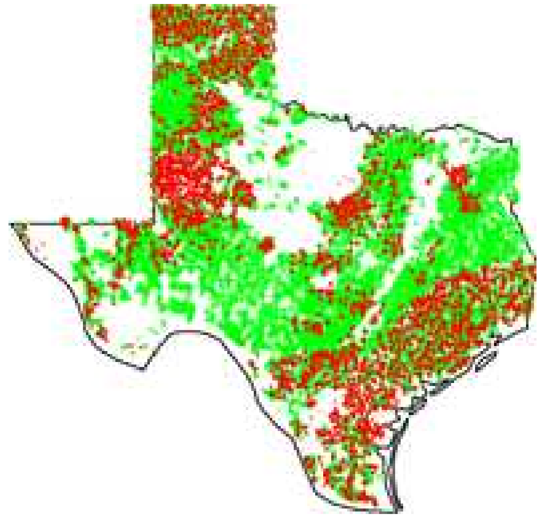

Figure 1: Arsenic concentration distribution in Texas. Legend: green (or light grey) star – safe wells; red (or dark grey) dot – dangerous wells.

senic concentration level to classify wells into two classes: *safe* and *dangerous*. A well whose arsenic concentration level is above $10\mu g/l$ is considered to be dangerous, based on the standard for drinking water by Environment Protection Agency [1]. Figure 1 illustrates the distribution of arsenic concentration in state Texas.

Continuous attributes excluding latitude and longitude are first discretized into categorical attributes based on the Recursive Minimal Entropy Partitioning technique [11]. For example, the concentration of nitrate has been discretized into seven intervals of (0-0.085], (0.085-0.125], (0.125-0.135], (0.135-0.265], (0.265-16.1], (16.1-28.31], (28.31-∞) (measurement unit mg/l). Notice that the interval is not in equal length. Then all the categorical attributes are transformed into binary attributes, which results in 250 items for each transaction.

The longitude and latitude of a well along with its arsenic class label are used to identify interesting regions, using the stratification techniques that we discussed in section 3.3. We found four interesting regions from the GWDB. As illustrated in Figure 2, Region 1 and Region 3 have high density of dangerous wells, and Region 2 and Region 4 have high density of safe wells. We applied supervised association rule mining in the whole Texas. We then stratified transactions by the four regions and mined regional rules as well.

After some exploratory experiments, we chose 0.01 for $min\_support$ and 0.7 for $min\_confidence$ for both global and regional rules mining. For our experiments, we observe that once $min\_support$ is lowered to 0.01, we will be able to find more interesting rules. The need to use low support values has also been observed by [16].

We compared the sets of rules generated for Region 1 and Region 3 (high density of dangerous wells), Region 2 and Region 4 (high density safe wells). The spatial risk patterns associated with arsenic are very different in each region. For

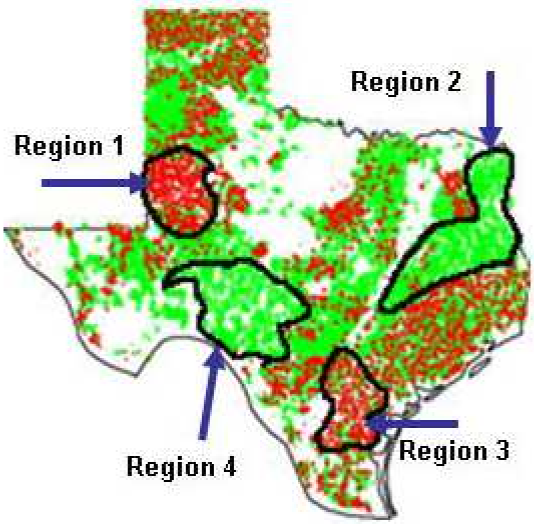

Figure 2: Interesting regions (bound by bold line) are identified.

example, the following is one of the interesting rules discovered in Region 1 (Figure 2):

$$
\begin{aligned}
&is\_a(X, Well) \wedge depth(X, 0 - 252) \wedge \\
&nitrate(X, 0.135 - 0.265) \wedge fluoride(X, 0.295 - 2.445) \\
&\rightarrow aresnic\_level(X,\ dangerous)\ \ (74\%).
\end{aligned} \tag{1}
$$

The rule states that with 74% confidence, wells in Region 1, which are up to 252 deep, with nitrate concentration between 0.135 and 0.265 mg/l and fluoride concentration between 0.295 and 2.445 mg/l, have dangerous arsenic concentration level.

However, a rule extracted from the Region 3 is:

$$
\begin{aligned}
&is\_a(X, Well) \wedge water\_use(X, \text{"}by\ humam\ beings\text{"}) \wedge \\
&nitrate(X, 28.31 - \infty) \wedge fluoride(X, 0.005 - 0.195) \\
&\rightarrow aresnic\_level(X,\ dangerous)\ \ (87\%).
\end{aligned} \tag{2}
$$

Instead of being related with well depth and relatively low concentration of nitrate (less than 0.265), the rule says that with 87% confidence, wells in Region 3, which are used by human beings, with nitrate concentration higher than 28.31 mg/l, and fluoride concentration between 0.005 and 0.195 mg/l, have dangerous arsenic concentration level.

Both regional rules (1) and (2) are not identified by the global rule mining at the global level. Statewide rule mining finds very general rules, for example:

$$
\begin{aligned}
&is\_a(X, Well) \wedge water\_use(X, \text{"}by\ humam\ beings\text{"}) \wedge \\
&arsenic\_level(X, safe) \\
&\rightarrow inside(X, Basin19)\ \ (86\%)
\end{aligned}
$$

It says that wells used by human beings, with safe arsenic concentration level are very likely (confidence is 86%) located in river basin 19. Our experiments showed that global patterns that are associated with dangerous wells cannot be discovered by the statewide mining. The generated regional rules seem to be more interesting for domain experts.

5. CONCLUSIONS

The paper introduced a novel framework to mine association rules relying on a given class structure. In contrast to traditional association rules, transaction are assumed to belong to a finite set of classes, and this class structure plays a key role for the preprocessing, discretization of continuous attributes, candidate itemset generation and pruning. We claim that this framework leads to a class-focused generation of association rules that sheds more light on the patterns related to a particular set of classes of interest. Consequently, this approach reduces the number of association rule generated significantly.

Moreover, a unique transaction stratification approach has been introduced that allows identifying interesting sub-regions in a spatial datasets for which regional association rules are then generated. A supervised clustering framework that identifies such regions has been discussed. Having low entropy with respect to the given class structure has been one of the guiding principle of many algorithms discussed in this paper.

We also conducted experiments that applied our techniques to a spatial datasets centering on discovery patterns of risk from arsenic based on Texas well data. Our approach was not only able to discover global rules with respect to the risks of arsenic as well as quite different regional rules, some of which are quite interesting for domain experts.

Our future work will center on applying our techniques to larger arsenic datasets that also include population, geological, and agricultural data. We also plan to generalize our framework to discover co-location patterns in spatial datasets. Finally, our novel stratification approach needs to be evaluated more thoroughly, and the efficiency of the employed algorithms has to be improved to be able to cope with very large spatial datasets.

6. ACKNOWLEDGEMENTS

We thank Dr. Shuhab Khan (Geosciences Department, University of Houston) for his valuable comments on experiments. We thank Radu Boghici and Roger M. Quincy (Texas Water Development Board) for help on Texas Ground Water Database. We thank Dr. Ping Chen (Computer Science Department, University of Houston-DownTown) for useful discussions.

7. REFERENCES

[1] U.S. Environmental Protection Agency. *http://www.epa.gov/*, 2006.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[4] Texas Water Development Board. *http://www.twdb.state.tx.us/home/index.asp*, 2006.

[5] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995.

[6] M. J. Egenhofer and R. D. Franzosa. Pointset topological spatial relations. *International Journal for Geographical Information Systems*, 5(2):161–174, 1991.

[7] Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, B13:238–241, 1951.

[8] Christoph Eick, Banafsheh Vaezian, Dan Jiang, and Jing Wang. Discovering of interesting regions in spatial data sets using supervised cluster, submitted. In *PKDD'06, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006.

[9] Smith A. H. et al. Cancer risks from arsenic in drinking water. In *Environmental Health Perspectives*, volume 97, pages 259–267, 1992.

[10] Carem Fabris and Alex Freitas. Discovering surprising instances of Simpson's paradox in hierarchical multidimensional data. *International Journal of Data Warehousing and Mining*, 2(1):26–48, January 2006.

[11] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Morgan Kaufmann, editor, *Proceedings of the 13th International Joint Conference on Artificial Intellogence*, pages 1022–1027, 1993.

[12] Clark Glymour, David Madigan, Daryl Pregibon, and Padhraic Smyth. Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1(1):11–28, 1997.

[13] Michael F. Goodchild. The fundamental laws of GIScience. Invited talk at University Consortium for Geographic Information Science, University of California, Santa Barbara, 2003.

[14] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering spatial co-location patterns from spatial datasets: A gneral approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12):1472–1485, December 2004.

[15] Krzysztof Koperski and Jiawei 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.

[16] Wenmin Li, Jiawei Han, and Jian 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.

[17] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pages 80–86, 1998.

[18] Mantel N. and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

[19] Salvatore Rinzivillo and Franco Turini. Extracting spatial association rules from spatial transactions. In *GIS '05: Proceedings of the 13th annual ACM international workshop on Geographic information systems*, pages 79–86, Bremen, Germany, November 4 2005.

[20] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 1–12, Montreal, Quebec, Canada, 4–6 1996.

[21] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[22] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *3rd SIAM International Conference on Data Mining (SDM'03)*, San Francisco, CA, May 2003.

[23] Jin Soung Yoo and Shashi Shekhar. A partial join approach for mining co-location patterns. In *GIS'04, Proceedings of the 12th annual ACM international workshop on Geographic information systems*, Washington, DC, November 12–13 2004.