Pre Prospection data mining model for hydrocarbon development guides to the need of developing a model for prospection activity for which the data of surface indicators collected through remote sensing and microbial sources can be used. This chapter proposes a new approach to mine context based positive and negative spatial association rules. Researchers are using Apriori algorithm on spatial databases which doesn’t utilize the strengths of positive and negative association rules hence ignoring discovery of very interesting and useful associations present in the data. In dense spatial databases, the number of negative association rules is much higher as compared to the positive rules which need exploitation. Positive and negative association rule discovery and then pruning uninteresting rules consumes lots of resources without much improvement in the overall accuracy in the knowledge discovery process. Overall analysis of spatial association rule mining concludes the absence of some unseen factors. The associations among different objects and lattices are strongly dependent upon the context where context is the state of entity, environment or action. We have proposed an approach of spatial association rule mining in which contextual situation is considered while generating positive and negative frequent item sets. An extended algorithm based on Apriori approach is developed and compared with existing spatial association rule algorithms. The numerical evaluation shows that proposed algorithm is efficient in generating specific, reliable and robust information than the traditional algorithms.

The increase in horizontal and vertical sizes of data in purpose built repositories makes databases an interesting laboratory for ascertaining hidden associations among various attributes that have been recorded for one specific reason or another. Scientists and engineers collect and work on terabytes of data received from space borne instruments and other remote sensing systems [Agarwal et al., 1993]. It is estimated that about 80% of corporate databases integrate spatial information [Shaheen et al., 2010] [Fayyad and Grinstein, 2001] along with other entities.
The detailed description of association rule mining/ spatio temporal association rule mining is given in Section 2.2.4.1.

4.1. Spatial / Spatio-Temporal Association Rule Mining

*Spatial data mining* is an expanding domain and it is widely accepted for the extraction of implicit knowledge, spatial relations and other patterns that are not explicitly stored in spatial databases. A *spatial association rule* is a rule which describes the implications of one set of features by another set of features in spatial databases [Koperski and Han, 1995]. In *spatial association rules*, spatial objects are related to each other by using spatial predicates and connectors [Vyas et al., 2007] such as Close to(), nearby(), adjacent_to(), etc. For example, it has been observed that 80% of ignited rocks that are closer to a river and nearer to vegetated fields contain *hydrocarbons*, as shown in Equation 4.1 where *Close_to*, *nearby* and *adjacent_to* are *spatial predicates*.

\[ \text{Close_to (A, river) } \land \text{nearby (A, vegetation) } \land \text{Is_a (A, ignited_rock)} \rightarrow \text{adjacent_to(A, potential_trap)} \] (4.1)

\text{Supp: 60\%, Confidence: 80\%}

The number of predicates in *consequent* and *antecedent* of the spatial association rule can vary depending upon the rule to be represented.

\[ \text{Nearby (B, ignited_rock)} \rightarrow \text{far_away (B, potential_trap)} \]

\text{Supp: 50\%, Confidence: 80\%} (4.2)

*Spatial association rules* can have both the spatial and non-spatial predicates whereas a spatial predicate is used to represent a spatial relationship which can have an association with non-spatial predicate as in Equation 4.3.

\[ \text{Nearby (A, ignited_rock)} \rightarrow \text{color_of(A, blue)} \]

\text{Supp: 60\%, Confidence: 100\%} (4.3)

These rules represent positive associations among different features of objects. Negative association rules can also be of potential interest, particularly in spatial association rule mining. A *negative association rule* can show, for example, that a potential trap has never been found adjacent to wet sand, as shown in Equation 4.4.

\[ \text{Nearby (A, wet_sand)} \rightarrow \neg \text{adjacent_to (A, potential_trap)} \]

\text{Supp: 50\%, Confidence: 90\%} (4.4)
Spatial association rule mining is an area of potential interest for researchers. Further details of spatial association rule mining can be found in [Koperski and Han, 1995; Sharma et al., 2005], [Bembinek and Rybinski, 2009; Shehkar et al., 2003; Bogorny et al., 2008]. Spatial databases can render the attributes to depict the existence of an object in space. Space and time are considered in tandem to describe the existence of and changing patterns of matter and energy. A spatio temporal process is a multi dimensional space in which spatial states, spatial and non spatial attributes and spatial patterns vary continually across time [Shaheen et al., 2010; Xuewu et al., 2008]. Such a process is represented by a set of states at different time intervals. The variation in space across time can reflect patterns in the process of evolution. The number of states in the spatio temporal process depends on the size of the time granularity and time step. A spatio temporal association rule is extracted from multiple instances of spatial databases for which each instance is recorded after a particular time interval. The selection of time granularity and time step is dependent upon the application domain. The figure below describes the process of spatio temporal association rule mining.

Figure 4.1. Spatio- temporal association rule mining
Spatio temporal association rule mining has wider acceptance in different applications like climate change pattern detection, hydrocarbon reservoir characterization, natural resource accumulation and urban feature mapping, etc. The literature on spatio temporal association rule mining has been reviewed by [Verhein and Chawla, 2006; Shu et al., 2008; Mennis and Wei Liu, 2005; Compeita et al., 2007; Calargan and Yazici, 2008].

4.2. Problem Definition

Consider a spatial database, SD, containing geographical spatial and non-spatial data of different sites hydrocarbon’s pipeline network assets over multiple time periods. Let $X = \{X_1, X_2, X_3, ..., X_n\}$ be the set of objects contained in the database which are spatially located on the satellite image. Let $S$ be the subset of items in $X$. Let $TI=\{TI_1, TI_2, TI_3, ..., TI_n\}$ are the time intervals on which the data is collected and $site=\{site_1, site_2, site_3, ..., site_n\}$ are the sites from which the data is collected. $TI_i$ and $Site_j$ represents $i^{th}$ and $j^{th}$ instances of time interval and site.

Let $context_m[k]$ be the array structure which stores the values of the influencing parameter, for example, $context$ where $m$ is the name of the $context$ variable and $k$ is the array subscript. The Apriori Algorithm of association rule mining [Laube et al., 2008] and modified Apriori algorithm for positive and negative association rule mining is based on the following.

For an association rule $X \rightarrow Y$

- The rule is considered valid only when the value of support satisfies
  - For positive association rules:
    \[ Supp (XUY) > ms \text{ (ms = minimum support)} \] (4.5)
  - For negative association rules:
    \[ Supp (X)>ms, Supp(Y)>ms \text{ and } Supp (XUY) <ms \] (4.6)

- Interestingness measure for both positive and negative association rule satisfies
  \[ |Supp (XUY) – Supp(X) Supp(Y)| > mi \text{ where (mi=minimum interestingness)} \] (4.7)

In a spatio temporal database, if both positive and negative association rules are considered, then the number of rules will become enormous. An interestingness measure is defined to select appropriate rules. The problem is that the rules that are not interesting are removed immediately. There is no mechanism to deal with those rules which are falsely generated at some $ti_i$ because of an abnormal value of $context_m$. The scenario is depicted in Figure 4.2.
In the absence of context variable, the situation depicted above will ignore the rule $R_3$. The rule $R_4$ at time interval $TI_2$ is similar to $R_3$ at time interval $TI_1$ and $T_3$ except for the difference in the color of the water. We can see that the context of $TI_2$ (i.e., temperature at $TI_2$) doesn’t lie within the normal range, hence affecting the color of the water. The accuracy of the association rule is affected by an influencing factor which is named, ‘context’. The rule $R_4$ is not there in the intersection set $TI_1 \cap TI_2 \cap TI_3$ whereas this would be a valid rule if the context does not exert influence. The same rule lies in the intersection set $TI_1 \cap TI_3$. The rule is ignored because it doesn’t meet the criteria of minimum support. This criterion should be redesigned both for positive and negative association rules on the basis of context as an influencing variable. The support-confidence framework is modified to accommodate context as an affecting factor and to mine both positive and negative association rules on the basis of a new support value.

Figure 4.2. Spatio-temporal association rule mining with different contexts
4.3. A Method For Mining Context Based Spatio Temporal Association Rules

We developed a context based method to mine both positive and negative association rules from spatial databases. The algorithm for mining positive and negative association rules has already been devised by Wu et al. [2004] and implemented on spatial databases by Sharma et al [2005]. The same method is extended here to mine association rules on the basis of context. Positive and negative association rules are incorporated in support-confidence framework on the basis of the following principles [2004].

A positive association rule \( A \Rightarrow B \) is of potential interest if

1. \( A \) and \( B \) are disjoint item sets i.e., \( A \cap B = \emptyset \)
2. \( \text{Supp}(A) > ms; \text{Supp}(B) > ms; \text{Supp}(A \cup B) > ms \)
3. \( \text{Supp}(A \Rightarrow B) = \text{Supp}(A \cup B) \)
4. \( \text{Conf}(A \Rightarrow B) = \frac{\text{Supp}(A \cup B)}{\text{Supp}(A)} \)

Similarly, the negative association rule \( A \Rightarrow \neg B \) is of potential interest if

1. \( A \) and \( B \) are disjoint item sets i.e. \( A \cap B = \emptyset \)
2. \( \text{Supp}(A) > ms; \text{Supp}(B) > ms; \text{Supp}(A \cup B) < ms \)
3. \( \text{Supp}(A \Rightarrow \neg B) = \text{Supp}(A \cup \neg B) \)
4. \( \text{Conf}(A \Rightarrow \neg B) = \frac{\text{Supp}(A \cup \neg B)}{\text{Supp}(A)} \)

This will result in an exponential number of association rules which can be pruned by using the interestingness measure, as given in (Equation 4.7.) [Wu et al., 2004].

\[
\text{Interest}(X,Y) = |\text{Supp}(X \cup Y) - \text{Supp}(X)\text{Supp}(Y)|
\]

The integration of the interestingness measure in a support-confidence framework for positive association rule mining will result in:

\[
fip(I) = \text{Supp}(I) > ms \land \exists X, Y : X \cup Y = I \land fipi(X,Y)
\]

Where

\[
fipi(X,Y) = X \cap Y = \phi \land f(X,Y,ms,mc,mi) = 1
\]

\[
f(X,Y,ms,mc,mi) = \frac{\text{Supp}(X \cup Y) + \text{Conf}(X \cup Y) + \text{Interest}(X \cup Y) - (ms + mc + mi)}{(\text{Supp}(X \cup Y) - ms) + (\text{Conf}(X \cup Y) - mc) + (\text{Interest}(X \cup Y) - mi)}
\]
Fipi represents frequent itemset of potential interest. ms is the user specified minimum support, mc is the user specified minimum confidence and mi is the user specified minimum interest.

The interestingness measure is also incorporated in the support-confidence framework for negative association rule mining in the following way below [Wu et al., 2004]. J is an infrequent itemset of potential interest (iipis) if:

\[
\text{iipis}(J) = \text{Supp}(J) < ms \land \exists X, Y : XUY = J \land \text{iipis}(X, Y) \\
\text{iipis}(X, Y) = X \cap Y = \phi \land g(X, \neg Y, ms, mi) = 2
\] (4.9)

\[
g(X, \neg Y, ms, mc, mi) = f(X, Y, ms, mc, mi) + \frac{\text{Supp}(X) \cdot \text{Supp}(Y) - 2ms + 1}{|\text{Supp}(X) - ms| + |\text{Supp}(Y) - ms| + 1}
\]

To incorporate context variable in support-confidence framework, we need to understand that context can be different for different sets of circumstances. The selection of context variable is application dependent with the assumption that the user will select any parameter as a context variable which holds a positive value. Once the context variable is selected, a valid range of values from initial_value_context_rerange to final_value_context_rerange will be specified for context_m. The selection of initial and final values for defining the range of context variable is also an expert’s job. The initial value of context range is the value which, if decremented, would bring a negative change in the frequency of an item’s occurrence. Similarly, the final value is the one which, if incremented, causes a positive change in the frequency of an item’s occurrence. As for example, temperature is considered to be a context variable in Figure 4.2. The valid range for a context variable is from 30 to 50 in this application. If the temperature lies within the range, no major change in the signature of the selected objects is observed. A decrease in the initial value of temperature (i.e., 30) will decrease the frequency of the occurrence of certain spatial objects. If the value of context_m lies within the range of the initial to final values, the support is calculated using the support method mentioned above. If the value of context_m doesn’t lie in the normal range, then there can be four possibilities—these are detailed in Sections 4.4.1 and 4.4.2. To summarize the above before going into the details of algorithm, the user will explicitly define:

1. What parameter is considered to be the context variable?
2. What is the normal range of the context variable?
3. The smallest value of a context variable range is the initial value, if any decrement in the value cause a negative change in the frequency of item’s occurrence.

4. The largest value of context variable range is the final value, if any increment in the value causes a positive change in the frequency of item’s occurrence.

5. The scenario demonstrated in points 3 and 4 will be inverted according to the change in frequency of an item’s occurrence.

4.3.1. Positive association rules

Once the normal range for the context variable is defined, there can be two possibilities for both positive and negative association rules. Either the value of the context variable lies in the normal range or vice versa. If the value lies in the normal range, then association rules are extracted using a simple positive and negative association rule mining algorithm as explained above and proposed in [Sharma et al., 2005]. In the other case, when the values do not lie in the normal defined range, there can be two more possibilities:

(1). The value of context variable will be greater than the upper limit of the provided range; or

(2). The value of context variable will be less than the lower limit of the provided range.

Both values indicate abnormality in context. When the value of context variable is greater than the upper limit of provided range, the co-occurrence of item with each other in itemset increases. This is compensated by subtracting that increase from the support value because the increase in the value of the context variable caused the support value to increase by a small amount. Compensation is made by subtracting the percent increase from support. In the second case, a decrement in the value of the context variable is compensated by adding a percent decrement in the support value. This is illustrated in the following four cases:

Case I: The value of context\(_m[k]\) is lower than initial_value_context\(_m\)range. In this case support is calculated by finding the percentage difference of context\(_m[k]\) and initial_value_context\(_m\)range and storing it in positive_difference_context\(_m\). The positive_difference_context\(_m\) percent value of support is calculated and added to the support value to get a new support value.

\[
\text{positive_difference_context}_m = \frac{(\text{initial_value_context}_m\text{range} - \text{context}_m[k]) \times 100}{\text{context}_m[k]}
\]

\[
\text{Supp } (X \Rightarrow Y) = \text{Supp } (XUY) + \left( \text{Supp}(XUY) \times \frac{\text{Positive_Difference_context}_m}{100} \right)
\]
Case II: The value of context$_m[k]$ is greater than the final_value_context$_m$range. The percentage difference of final_value_context$_m$range and context$_m[k]$ are then calculated and the resulting percentage difference of the old support value is calculated. The difference is subtracted from the old support value to get a new support value. The variable negative_difference_context$_m$ is used to store the value of the difference.

\[
\text{negative\_difference\_context}_m = (\text{context}_m[k] - \text{final\_value\_context}_m\text{range}) \times 100 / \text{context}_m[k] \tag{4.12}
\]

\[
\text{Supp (X=>Y)} = \text{Supp (XUY)} - (\text{Supp(XUY)} \times \text{Negative\_Difference\_context}_m/100) \tag{4.13}
\]

4.3.2. Negative association rules

Case III: For negative association rules, positive_difference_context$_m$ is calculated by using the same equation (i.e., Equation 4.10). The positive_difference_context$_m$ value is then subtracted from the original support because negative association rule ensure non co-occurrence of the items together with an underlying assumption that the support values for both individual items of itemset are greater than the specified threshold.

\[
\text{Supp (X=>Y)} = \text{Supp (XUY)} - (\text{Supp(XUY)} \times \text{Positive\_Difference\_context}_m/100) \tag{4.14}
\]

Case IV: Similarly, the negative_differnce_context$_m$ percent value is added to the normal support value to eliminate the effect of contextual change.

\[
\text{Supp (X=>Y)} = \text{Supp (XUY)} + (\text{Supp(XUY)} \times \text{Negative\_Difference\_context}_m/100) \tag{4.15}
\]

In the view of above, we derived that the support of the spatial association rule is influenced by context. Hence, the value of support is compensated by the percentage difference in order to remove the effects of context. It is obvious that the addition and subtraction of a context variable in support value may not provide full compensation but most of the rules which were eliminated because of an abnormal value of context variable are observed to be part of this solution. Similarly, 40% of the rules which were falsely included with previous support are not part of this solution. After calculating the support value, uninteresting rules are pruned using the interestingness measure devised by Wu et al. [2004], as in Equation 4.7.


4.4. Algorithm

Name: Context based positive and negative spatio temporal association rule mining

**Inputs:** SD: spatial database;
Icontext<sub>m</sub>: current value of context variable;
IVCR: Initial value for context variable range;
FVCR: Final value for context variable range;
TI: Time intervals;
ms: minimum support;
mc: minimum confidence; and
mi: minimum interestingness.

**Outputs:** FIS<sub>p</sub>: Positive Frequent Item sets; IFIS<sub>n</sub>: Negative Infrequent Item sets;

1. Let FIS<sub>p</sub> = 0; IFIS<sub>n</sub> = 0; //Initialize FIS<sub>p</sub> and IFIS<sub>n</sub>
2. Set IVCR = input<sub>1</sub>; Set FVCR = input<sub>2</sub> // Initialize the initial and final values of context variable range
3. For (i=1; i<n; i++) // Data is extracted at multiple time intervals from 1..n
   Begin
   L<sub>i</sub> = get_predicates(SD, TI<sub>i</sub>); //predicates at multiple time intervals TI<sub>i</sub>
   FIS<sub>p</sub> = FIS<sub>p</sub> U L<sub>i</sub>;
   End
4. For(j=2; (L<sub>j-1</sub> ≠ Ø); j++) do
   Begin
   A<sub>k</sub> = get_candidate_set (L<sub>j-1</sub>); //candidate itemsets
   For each object s in S do begin
   A<sub>s</sub> = get_subsets(A, s);
   For each itemset A in A<sub>s</sub> do
   A.count = A.count + 1;
   End
   L<sub>j</sub> = {c | cєA<sub>k</sub> ∧ supp_context (c, IVCR, FVCR, Icontext<sub>m</sub>) >= ms};
   N<sub>j</sub> = A<sub>k</sub> - L<sub>j</sub>;
   // Pruning of uninteresting itemsets in L<sub>j</sub>
   For each itemset m in L<sub>j</sub> do

If NOT (fipi (m)) then
L_j = L_j – \{m\};
FIS_p = FIS_p \cup L_j;

// Pruning of uninteresting itemsets in N_j
(8). For each itemset n in N_j do
   If NOT (iipi(n)) then
      If (supp_context(c,IVCR,FVCR,lcontextm))<ms
          N_j = N_j – \{n\};
          IFIS_n = IFIS_n \cup N_j;
   End
(9). Output FIS_p, IFIS_n

Supp_context (r, initial_value_contextm,range, final_value_contextm,range, context_m)
Begin
R = r.supp/ SD; // Support count
if (context_m < initial_value_contextm,range) then
   begin
   For (Positive_rules)
      positive_difference_contextm = (initial_value_contextm,range – context_m)*100/ context_m
      R = r + (r*Positive_Difference_contextm,100/100) // new support value
   End
   Elseif(context_m> final_value_contextm,range, context_m) then
   Begin
   Negative_Difference_contextm = (context_m -final_value_contextm,range)*100/context_m
   R = r + (r*Negative_Difference_contextm,100)
   End
   For (Negative_rules)
      positive_difference_contextm = (initial_value_contextm,range – context_m)*100/ context_m
      R = r - (r*Positive_Difference_contextm,100) // new support value
   End
   Elseif(context_m> final_value_contextm,range, context_m) then
   Begin
   Negative_Difference_contextm = (context_m -final_value_contextm,range)*100/context_m
   End
End
$R = r - (r*\text{Negative\_Difference\_context}_m/100)$

End

Return

The procedure discovers all frequent and infrequent item sets from the spatial database on the basis of a context variable. Users may apply a Context variable at their discretion and this can be changed depending upon the need for rule mining and the given set of circumstances. In the first step of the proposed algorithm, $FIS_p$ and $IFIS_n$ are initialized. In step (2), the range for context variable is set which spans from the initial value to the final value. In step (3), predicates from a spatial database are extracted and stored in order of time intervals. In step (4), candidate sets and subsets are obtained as per the Apriori algorithm. In step (5), the function of support value calculation is called to produce frequent item sets and in step (6), infrequent item sets are produced. In steps (7) and (8), uninteresting item sets from both the frequent and infrequent item sets are eliminated. The output is produced in the final step (9). The procedure is illustrated in the following flowchart.
Figure 4.3. Flowchart for mining context based positive and negative spatio temporal association rules
Two conditional controls are used in the above flowchart. The value of the context variable can lie in two ranges. If the value lies in the normal range, the conventional procedure of positive and negative association rule mining is used. In the case of an abnormal value, either the value will be greater than the final value of range or it will be less than initial value of the range. Both cases are dealt with according to the procedure given in sections 4.4.1 and 4.4.2.

4.5. Experimentation

The algorithm is implemented on multiple thematic maps of different times which were collected from the hydrocarbon pipeline industry and exploration companies. The pipeline network is mapped onto a geographic information system (GIS). Using cartographic mapping, rectification is performed in ESRI’s ArcGIS 9.2 whereas the spatial database is managed in Microsoft SQL Server. The thematic maps consist of multiple layers including pipeline, sales metering stations, cathodic protection stations, town border stations, reducers, buried and over ground valves, and different types of consumers and parcels. The GIS designed for exploration purposes also contains multiple layers including rocks, vegetation, soil, water, concrete structures, land, shadows, residential areas, roads and topography. All these layers are stored at the backend in the form of individual tables along with spatial tables. The details of layers and pertaining spatial data of both GIS are detailed in Tables 4-1 and 4-2.

Table 4-1. Spatial and Non-Spatial Data of Pipeline Network

<table>
<thead>
<tr>
<th>Sr #</th>
<th>Name of Layer</th>
<th>No of attributes</th>
<th>No of non-spatial records</th>
<th>No of records on the basis of spatial combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sales metering stations</td>
<td>12</td>
<td>3456</td>
<td>Spatial combination operators: adjacent_to, nearby, far_away, close_to, parallel_to</td>
</tr>
<tr>
<td>2.</td>
<td>Town border stations</td>
<td>19</td>
<td>4850</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Cathodic protections stations</td>
<td>17</td>
<td>8440</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Pipeline</td>
<td>10</td>
<td>54462</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Buried valves</td>
<td>14</td>
<td>9450</td>
<td></td>
</tr>
<tr>
<td>No. of records: 5,81,504</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of intervals:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(TI [n] = 12)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of records: 6124</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of intervals:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(TI [n] = 7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-2. Spatial and Non-Spatial Data of Exploration GIS

<table>
<thead>
<tr>
<th>Sr #</th>
<th>Name of Layer</th>
<th>No of attributes</th>
<th>No of non-spatial records</th>
<th>No of records on the basis of spatial combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Rocks</td>
<td>10</td>
<td>860</td>
<td>Spatial combination operators: adjacent_to, nearby, far_away, close_to, parallel_to</td>
</tr>
<tr>
<td>2.</td>
<td>Vegetation</td>
<td>7</td>
<td>820</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Soil</td>
<td>21</td>
<td>852</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Water</td>
<td>12</td>
<td>848</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Concrete structures</td>
<td>8</td>
<td>810</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Land</td>
<td>22</td>
<td>880</td>
<td>No. of records: 6124</td>
</tr>
<tr>
<td>7.</td>
<td>Shadows</td>
<td>6</td>
<td>800</td>
<td>No. of intervals:</td>
</tr>
<tr>
<td>8.</td>
<td>Residential area</td>
<td>10</td>
<td>850</td>
<td>(TI [n] = 7)</td>
</tr>
<tr>
<td>9.</td>
<td>Topography</td>
<td>6</td>
<td>842</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Roads</td>
<td>8</td>
<td>833</td>
<td></td>
</tr>
</tbody>
</table>
The rational of this experiment is to compare the performance of proposed algorithm with two commonly used algorithms for association rule mining. The performance of the association rule mining algorithm can be measured on the basis of the number of extracted rules. In the case of spatial data, the number of association rules is sometimes misleading and undesirable hence the performance might be compared on the basis of the rules and against the quality of the association rules. The confidence measure can help predict the quality of the rule. The execution time of the algorithms is also compared to the rank of its usability.

The proposed algorithm is specifically designed for spatio temporal data for which various satellite images collected over a certain period of time may better reflect the execution performance of the algorithm.

The records summarized in Tables 4-1 and 4-2 are mapped to a spatial database having spatial attributes such as adjacent_to, close_to, far away, nearby, etc. and non-spatial attributes such as length_of_groundbed_CP, size_of_TBS_manifold, lubrication_condition_valves, total_no_of_CPs, etc. A total number of 581504 spatial records are oriented at 12 different time intervals from the period 1990 to 2010 in pipeline GIS and 6124 records of exploration GIS. After object recognition from a satellite image, the spatial data are converted to show spatial relationships among different objects. The predicate clauses are decided on the basis of distance, geometry and shape of object.

Satellite images of a hydrocarbon reserve and its classified image are shown in Figure 4.4 and 4.5. Another image mapped with spatial predicates is shown in Figure 4.6.

The spatial records are stored in separate schema of the database with respect to the agreed upon predicates. These images are then classified to recognize different objects and features on the satellite image. The schema, having predicate based records, also contains the occurrence of different objects on the image.

The performance of algorithm is compared with two most commonly used algorithms i.e. Apriori Algorithm and the algorithm to mine positive and negative association rules [Wu et al., 2004]. All these performances are compared with respect to the following:

a. No. of spatio temporal association rules
b. Execution time
c. Confidence of rules
Figure 4.4. Remotely sensed image of prospect site

Figure 4.5. Classified image of prospect site
4.6. Results

We study the results of our algorithm with respect to the number of rules, average confidence of rules and total time taken to extract association rules (execution time). The extraction of association rules on the basis of context is a novel approach and has its own impediments. That is why the comparison may not manifest the significance of the approach. The comparison is in fact made to compare the use of an existing technique on spatio temporal data to the proposed technique. These results are compared with the results produced by Apriori and Positive/Negative Association rule mining algorithm.
Figure 4.7. No of rules Plot of Apriori, Positive/Negative ARM and Context based ARM (Proposed algo)

In Figure 4.7, the three algorithms are compared on the basis of number of rules. The Apriori algorithm always produces fewer rules than other algorithms. Three cases of the proposed algorithm are considered in this plot (1). If the value of the context variable (cv) becomes less than the initial value of the range (ci) (2). If the value of cv is greater than that of the final value of the
context range (cf) (3). If the value of the cv lies between the normal range. The algorithm produced
a maximum number of rules with assumption (1) up to the 0.5 value of minimum support, whereas
it produced comparatively fewer rules with the assumption (2) because, in that case, the value of
the context variable exceeds the upper limit of the range. The insight brought by the plot of figure
4.7 can be evaluated against different aspects. Keeping other performance measures in view, the
greater the number of association rules, the greater the patterns extracted from the database.
Increase in support value causes the algorithm’s results to show a divergence towards a single
point because there are very few co-occurrences in real datasets with much larger support. The
Apriori algorithm produced the fewest rules because it ignores negative association rules. The
results of PNARM lie almost in between two cases of CBPNARM. The average number of rules in
PNARM seems to be greater than number of rules extracted from CBPNARM but the significance
of the CBPNARM algorithm can be viewed by analyzing both the plots of Figure 4.7 and 4.8.

In the plot (Figure 4.8), the confidence of rules extracted using our algorithm has the
highest projection of all at various support values. The greater the value of confidence, the greater
the accuracy of rules with the exception of the dependence of the variables. The confidence of
rules gives a true projection up to a certain value of support because the increase in support returns
very few co-occurrences from the data, hence limiting a broader capability to evaluate the
algorithm. CBPNARM produced rules with greater confidence up to 0.5 support even when
compared with the Apriori algorithm. The rules produced by PNARM are at the lowest confidence
which demonstrates the significance of considering contextual information, especially for positive
and negative spatial association rule mining—as far as PNARM produced the largest number of
rules but with a smaller confidence value.

The algorithm is an extended form of PNARM inheriting its total capability with an
additional capability of context based mining. The execution time of the proposed algorithm is a
bit higher as per our expectations. The additional module takes some extra time which seems to be
negligible with the increase in minimum support. In Figure 4.9, the three algorithms are compared
on the basis of execution time.
Association rule mining is considered to be one of the most widely used data mining techniques. It is likely to be used for extracting associations among large spatial datasets applied in natural resource exploration, network management, geological mapping and crime analysis, etc. The algorithm presented in this study is an extended form of the algorithm presented for positive and negative association rule mining. A new factor, ‘context’, is introduced in this research which is important to consider in mining association rules. The proposed algorithm offers a method for dealing with context, which explicitly affects the accuracy of association rules. The selection of context variable(s) and its range depends on user input which should be executed carefully. For implementing the algorithm, spatial data is collected at multiple instances of time and structured systematically on the basis of context. The algorithm extracts the rules from a large spatial database at the given value of support and confidence.

Our evaluation of the algorithm shows that it is more accurate in terms of its granularity of output rules and confidence. The granularity of output rules is consistent and the rules are produced with greater confidence. Though the execution time of the algorithm is greater than that of previous algorithms, the extracted patterns are decision-oriented, specific and clear. The increase in execution time is due to the inclusion of an external factor (i.e., a context which is not part of the adopted procedural approach and can be considered as an external influencing factor).

As already discussed that there are two data sources that will be used for prospecting hydrocarbon. (1). Remote sensing (2). Microbial indicators. This chapter focused on the first one.
where the used data was spatio temporal in nature. In the next chapter another method for prospecting hydrocarbon is proposed that uses microbial data. Microbial database uses general numeric and categorical data which with another angle of view validates the CBPNARM algorithm proposed in this chapter.