

Striking Two Birds With One Stone: Simultaneous Mining of Positive and Negative Spatial Patterns

Bavani Arunasalam *

Sanjay Chawla[†]

Pei Sun[‡]

Abstract

We propose an efficient algorithm to mine positive and negative patterns in large spatial databases. The algorithm is based on exploiting a complementarity property for a certain support-like measure. This property guarantees that if a positive k -pattern is "frequent" then $O(k)$ related negative patterns will be infrequent. For the traditional support measure this complementarity property holds true only when the minimum support is over fifty percent. We also confirm the correctness of our approach using Ripley's K-Function, a standard tool in spatial statistics for analyzing point patterns. Extensive experimentation on data extracted from the Sloan Digital Sky Survey (SDSS) database demonstrates the utility of our approach to large scale data exploration.

1 Introduction.

Spatial data mining is the task of discovering, extracting and learning patterns (relationships) in spatial databases. Spatial data is fundamentally different from transactional data in its fundamental nature and distributional tendencies. The objects in a spatial database are characterized by a spatial location and several non-spatial attributes. For example, a galaxy database may contain the x , y and z co-ordinates of the galaxies, the types of galaxies and several other attributes. An example of a spatial data mining task here would be to determine the confidence of finding a *spiral* galaxy in the neighbourhood of an *elliptic galaxy*.

In terms of spatial statistics this problem could be stated as follows: Given a set of points $S = \{x_i\}$, is there a way to quantify that the points in S exhibit a *random*, *correlated* or *negatively correlated* behavior? Scientists working in diverse fields including ecology, geology and astronomy are interested in inferring such information from their data sets as it often provides insights about the underlying mechanisms at play.

The standard tool to test for such patterns is the two-point correlation function ζ . Intuitively ζ calculates the likelihood for finding a point near the vicinity of another given point and can be expressed as

$$\zeta(\delta A) = \frac{N_{obs}(\delta A)}{N_{background}(\delta A)} - 1$$

where $N_{obs}(\delta A)$ is the number of points observed in a small area δA and $N_{background}(\delta A)$ is the expected number of points that will be observed assuming that the data is sampled from a given background distribution. Now the test to characterize the data set can be summarized as

$$\zeta = \begin{cases} < 0 & \text{then S is negatively correlated} \\ \approx 0 & \text{then S agrees with background} \\ > 0 & \text{then S is positively correlated} \end{cases}$$

However choosing a background model is a non-trivial exercise and is largely domain-dependent. Furthermore is the zeta function independent (homogeneous) or dependent upon the coordinates (inhomogeneous) or is it only dependent on the distance between points (isotropic) or does the direction matter (an-isotropic)? Depending upon the assumption, various ζ functions have been proposed in the literature.

Our objective is to use methods from spatial association rule mining to discover complex spatial relationships and then use statistical techniques to determine if these relationships are indeed substantive.

Complex relationships include both positive and negative association rules. Positive rules are the traditional association rules of the form $A \rightarrow B$, which indicates the presence of B in the neighbourhood of A . Negative rules are of the form $A \rightarrow -B$, which indicates the absence of B in the neighbourhood of A .

While much research has been done on positive association rules mining, very little progress has been achieved in mining negative patterns. The reason for this is while generating candidate itemsets in association rules mining, it would be necessary to consider not only the positive patterns, but also the negative patterns and each positive pattern of length k gives rise to $O(k)$ negative patterns making the search space exponentially

*University of Sydney, School of Information Technologies Sydney, NSW, Australia.

[†]University of Sydney, School of Information Technologies Sydney, NSW, Australia.

[‡]University of Sydney, School of Information Technologies Sydney, NSW, Australia.

larger than the space for positive patterns. In this paper we present an efficient algorithm to generate complex relationships in spatial databases.

1.1 Related work Association rules are considered one of the major success stories of data mining research [1]. Association Rules are traditionally described in the framework of market basket analysis. Given a set of items I and a set of transactions T consisting of subsets of I , an Association Rule is a relationship of the form $A \rightarrow^{s,c} B$, where A and B are subsets of I while s and c are the minimum support and confidence of the rule. A is called the antecedent and B the consequent of the rule. The support $\sigma(A)$ of a subset A of I is defined as the percentage of transactions which contain A and the confidence of a rule $A \rightarrow B$ is $\frac{\sigma(A \cup B)}{\sigma(A)}$. Most algorithms for association rule discovery take advantage of the anti-monotonicity property exhibited by the support level: If $A \subset B$ then $\sigma(A) \geq \sigma(B)$.

Our focus is to apply the principle of Apriori in spatial data. Koperski and Han [6] proposed the first extension of the Apriori paradigm to spatial data. However in their method they materialized all the possible spatial relationships that they intended to mine. This is equivalent to determining the universe of candidate interesting relationships. Thus in some ways their technique was hypothesis driven rather than hypothesis generating.

An efficient algorithm to mine a kind of spatial co-locations was proposed by Shekhar and Huang [9]. The concepts of neighbourhood, participation ratio, participation index were defined. Instead of support, the participation index was used as a pruning measure in the conventional Apriori-like technique. The drawback of their method is that some confident co-location rules with low support are also pruned. In order to solve this problem, Huang et al. [5] proposed the concept of maximal participation index and it was used as pruning measure to replace participation index. We will discuss these measures in detail in section 2.2, as they are central to our approach.

An algorithm to mine both positive and negative association rules was proposed by Wu et al. [11]. In their algorithm negative rules are generated from infrequent itemsets and interest is used as a further pruning measure. By considering the negative rules only in the infrequent itemsets, some potential confident negative rules could be lost. For example consider $supp(A)=0.5$, $supp(A,B)=0.25$ and $minsup=0.2$. Here $\{A,B\}$ is a frequent pattern since $supp(A,B) > minsup$. Therefore the negative pattern $\{A,-B\}$ will not be considered at all. However, $supp(A,-B) = supp(A) - supp(A,B) = 0.25$. This shows that even though $\{A,-B\}$ is a frequent pattern,

it is pruned in this approach. Therefore this method will not be able to mine the complete set of frequent negative patterns in the dataset.

Another approach for mining positive and negative relationships was proposed by Antonie and Zaiane [2]. In addition to the *minsup* and *minconf* measures they introduced the correlation threshold as the third parameter. Their approach generates only a subset of negative rules, which they refer to as *confined negative association rules*, where the entire antecedent or consequent must be a conjunction of negated attributes or a conjunction of non-negated attributes.

1.2 Key Insight and Main Contributions

1. Our objective is to simultaneously mine complex spatial relationships which include positive and negative patterns. We can achieve this objective using the following observation:

Suppose $\{A, B\}$ is a positive 2-itemset and $\sigma(A, B)$ its support. Then the following relationship can easily be derived

$$(1.1) \quad \sigma(A, -B) = \sigma(A) - \sigma(A, B)$$

Here $-B$ represents the absence of B . Now, because of the anti-monotonicity of the support measure, $\sigma(A) \geq \sigma(A, B)$. Therefore $\sigma(A, B) > 50\%$ implies that $\sigma(A, -B) < 50\%$. Note, a support greater than 50% is required for this to hold uniformly. This will enable the pruning of negative patterns (those that contain negative items) based on frequent positive patterns.

However, choosing a support level of greater than fifty percent will not lead to *interesting* results because either there will be very few itemsets which have such high support or the itemsets which do have such high support will probably be well known. Thus we need a measure M such that

- (a) A relationship similar to Equation 1.1 holds.
- (b) It is natural to set high threshold values (like confidence).
- (c) A monotonic-like property holds which will enable aggressive pruning of candidate patterns.

We will use the Maximal Participation Index (MaxPI) introduced by Huang et al. [5] and show that it satisfies all the three properties enumerated above.

To the best of our knowledge this is the first efficient algorithm which can simultaneously mine positive and negative patterns.

- In our earlier work [7], we had introduced a taxonomy of complex spatial relationships that are worthy of being mined. Table 1 gives the basic definitions of these relationships and we will show how to efficiently mine these relationships using our approach.
- We have carried out detailed experiments on a large extract of the Sloan Digital Sky Survey database to show how our approach can be used in practice. We discover patterns, which are known to be genuine and others which may turn out to be "interesting" thus confirming the role of data mining as a tool for credible hypothesis generation.

Relationship	Notation	Description	Example
Positive	$A \rightarrow B$	Presence of B in the neighbourhood of A	Sa type Spiral Galaxies \rightarrow Sb type Spiral Galaxies
Negative	$A \rightarrow -B$	Absence of B in the neighbourhood of A	Elliptic galaxies tend to exclude spiral galaxies. $E \rightarrow -S$.
Self-Co-location	$A \rightarrow A+$	Presence of many instances of the same feature in a given neighbourhood	Elliptic galaxies tend to cluster more strongly. $E \rightarrow E+$.
Complex	$A+ \rightarrow -C, B$	combination of two or more of the above relationships	Clusters of elliptic galaxies tend to exclude other types of galaxies. $E+ \rightarrow -S$.

- One of the weaknesses of association rule mining is that the number of rules that are generated far exceeds the "genuine" number of patterns that are interesting. We follow an approach analogous to the filter-refine strategy popular in spatial databases. We will use the NP_MaxPI algorithm to generate candidate patterns and use the Ripley's K-function to test and filter out rules, which are not substantive. However, our approach is more general compared to Ripley's K-function as Ripley's K-function can only find relationships between two features, whereas our approach can be applied to any number of features. Also Ripley's K-function finds relationships globally from the entire set of spatial objects whereas, our approach generates relationships locally within the neighbourhood of spatial objects. We will discuss Ripley's K-function in detail in section 4.3.

The remainder of the paper is organized as follows: Section 2 gives the definitions and the basic concepts involved in spatial data mining. In this section we also describe the Maximal Participation Index with examples, which is central to our approach. In Section 3 we introduce our approach to mining complex patterns including positive and negative rules. In this section we present a lemma, by using which our algorithm prunes a large number of negative patterns and hence effectively reduces the number of candidate itemsets. Section 4 presents our experimental results, which show the scalability, applicability and correctness of our algorithm. Section 5 concludes the paper with a summary of the research.

2 Basic Definitions and Concepts

In this section we present the basic concepts and definitions used in this paper.

Table 1: Types of Relationships

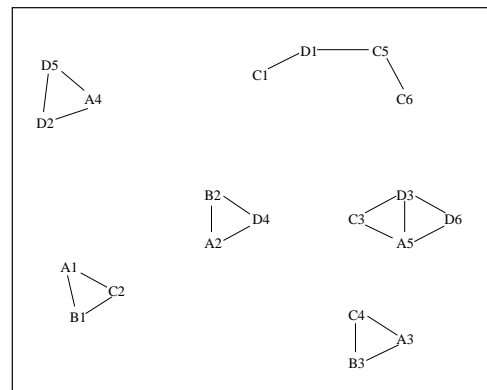


Figure 1: An Example of spatial co-location patterns

2.1 Co-location An important task in spatial data mining is extracting relationships between features that co-locate.

DEFINITION 2.1. (Co-location) Two spatial objects are said to co-locate if the Euclidean distance between the objects is less than or equal to the user-specified neighbourhood distance d .

Figure 1 gives an example of a set of spatial objects. In Figure 1, A_1 represents the ID of a spatial object of feature A, B_1 represents the ID of a spatial object of feature B and so on. The line joining two points indicates that the distance between them is less than or equal to the neighbourhood distance 'd' and hence they co-locate.

DEFINITION 2.2. (Clique) A set of spatial objects S is said to be a clique if every object in S co-locate with every other object in S and no super set of S has this property.

No	Clique
i	C_1, D_1
ii	C_5, D_1
iii	C_5, C_6
iv	A_4, D_2, D_5
v	A_5, C_3, D_3
vi	A_5, D_3, D_6
vii	A_1, B_1, C_2
viii	A_2, B_2, D_4
ix	A_3, B_3, C_4

Table 2: Cliques in Figure 1

Table 2 gives the cliques found in Figure 1.

Co-locational patterns are the relationships between the features in cliques of spatial databases. For example, in Figure 1 and Table 2, the co-locational pattern $\{A, C\}$ occurs 3 times, in v, vii and ix. Spatial relationships are confident co-locational patterns found in spatial databases.

2.2 Maximal Participation Ratio In this section we will briefly describe the notion of Maximal Participation Index (maxPI) as described by Huang et al. in [5] where more details can be found.

DEFINITION 2.3. (Participation ratio) Given a co-location pattern L and a feature $f \in L$, the participation ratio of f , $pr(L, f)$, can be defined as the support of L divided by the support of f . For example, in Figure 1, the support of $\{A, B, C\}$ is 2 and the support of C is 6, so $pr(\{A, B, C\}, C) = 2/6$.

DEFINITION 2.4. (Maximal Participation Index) Given a co-location pattern L , the maximal participation index of L , $maxPI(L)$ can be defined as the maximal participation ratio of all the features in L , i.e. $maxPI(L) = max_{f \in L} \{pr(L, f)\}$. For example, in Figure 1, $maxPI(\{A, B, C\}) = max(\frac{2}{6}, \frac{2}{3}, \frac{2}{6}) = \frac{2}{3}$.

A high maximal participation index indicates that at least one spatial feature (which we call the *maxfeature*) strongly implies the pattern. By using maxPI, rules with low frequency but high confidence can be found, which would otherwise be pruned by a support threshold [5].

2.2.1 Mining rules with low support and high confidence using maxPI As discussed above the maxPI measure could be used to generate rules with low support and high confidence. For example if A is an infrequent feature the support of $\{A, B\}$ will be very low and hence will be pruned by the support threshold.

However the confidence of the rule $A \rightarrow B$, defined as, $conf(A \rightarrow B) = \frac{support(A, B)}{support(A)}$ could be high.

This problem can be directly addressed by using the *maxPI* measure. $maxPI(A, B)$ is given as,

$$maxPI\{A, B\} = max(pr(\{A, B\}, A), pr(\{A, B\}, B)) \\ = max(conf(A \rightarrow B), conf(B \rightarrow A))$$

This shows that a high confidence for the rule $A \rightarrow B$ will lead to high maxPI, which will prevent rules with low support and high confidence from being pruned¹.

2.2.2 The weak monotonic property of maxPI

Maximal participation index is not monotonic with respect to the pattern containment relations. For example, in Figure 1, $(maxPI(\{A, C\}) = 3/5 < (maxPI(\{A, B, C\}) = 2/3$. Interestingly, as pointed out in [5] the maximal participation index does have the following weak monotonic property:

If P is a k -co-location pattern, then there exists at most one $(k-1)$ subpatterns P' of P such that $maxPI(P') < maxPI(P)$.

Relying on this weak monotonic property, the Apriori-like algorithm can be modified to mine confident patterns by using a maxPI threshold.

3 The NP_MaxPI Algorithm

The *maxPI* measure was introduced to discover *positive* co-location patterns with high confidence and low support. However our goal is to mine complex relationships which include both positive and negative patterns. We now show that *maxPI* is a good candidate to help achieve our goal, i.e., simultaneously discover both positive and negative patterns.

The main challenge in mining complex relationships is the high processing cost due the large number of candidate itemsets which include positive and negative features. Each positive candidate k -pattern will give rise to $O(k)$ k -patterns which contain a negative feature.

Let $F = \{A, B, C, D\}$ be the set of all features in the spatial database. Then for mining complex relationships, the candidate 1- itemsets would be $\{A, B, C, D, -A, -B, -C, -D\}$. Hence as the number of features in the spatial database increases, the candidate 1-itemsets is doubled. This would result in an exponential growth in the candidate space for larger itemsets. In this paper we propose an approach, which effectively reduces the number of candidate itemsets when mining for positive and

¹This is a strength and weakness of the maxPI measure. While patterns with low support and high confidence are likely to be "interesting", it is possible that they are an artifact of the data and may not be statistically significant.

negative patterns. Our approach is based on Lemma 1 which shows how a large number of negative patterns can be pruned when a positive pattern is greater than the threshold t where $t \geq 0.5$. Since $\max\text{PI}$ is based on the confidence, such a high threshold is not unusual for $\max\text{PI}$.

For example, let $F = \{A, B, C, D\}$ be a 4-itemset. Then, if $\max\text{PI}(\{A, B, C, D\}) = \text{pr}(\{A, B, C, D\}, A) > t > 0.5$, then we prove that $\max\text{PI}(\{A, -B, C, D\})$, $\max\text{PI}(\{A, B, -C, D\})$ and $\max\text{PI}(\{A, B, C, -D\})$ are all less than t and hence could be pruned.

Notations used:

1. Let $F_k = \{f_1, f_2, \dots, f_k\}$ be a k -pattern and $F_{k-l} = \{f_1, \dots, f_{l-1}, -f_l, f_{l+1}, \dots, f_k\}$.
For example, if $F_3 = \{A, B, C\}$ where $f_1 = A$, $f_2 = B$ and $f_3 = C$, then $F_{3-2} = \{A, -B, C\}$.
2. Let $\sigma(F_k) = \text{support}(F_k)$ and $\sigma(F_{k-l}) = \text{support}(F_{k-l})$.

LEMMA 3.1. Let F_k be a k -pattern and $t \geq 0.5$. Furthermore, assume that $\sigma(-f) > \sigma(f) \forall f$. If $\max\text{PI}(F_k) \geq t$ and $\max\text{PI}(F_k) = \text{pr}(F_k, f_j)$ then $\max\text{PI}(F_{k-l}) < t$ for every $f_l \in F_k, f_l \neq f_j$.

Proof. By definition, $\max\text{PI}(F_{k-l})$

$$= \text{Max} \left\{ \text{Max}_{i=1, i \neq l}^k \left\{ \frac{\sigma(F_{k-l})}{\sigma(f_i)} \right\}, \frac{\sigma(F_{k-l})}{\sigma(-f_l)} \right\}$$

In spatial data, the number of instances of absence of a feature f_l , i.e., $\sigma(-f_l)$, will be equal to the number of points in the area of observation which do not contain that feature. In real spatial datasets it is appropriate to assume that the support of absence of features would be greater than the support of presence of features.² Therefore,

$$\max\text{PI}(F_{k-l}) = \text{Max}_{i=1, i \neq l}^k \left\{ \frac{\sigma(F_{k-l})}{\sigma(f_i)} \right\}$$

But from the assumption,

$$\max\text{PI}(F_k) = \text{pr}(F_k, f_j) = \frac{\sigma(F_k)}{\sigma(f_j)}$$

This implies that $\sigma(f_j) \leq \sigma(f_i)$ for every $f_i \in F_k, i \neq j$. Hence,

$$\max\text{PI}(F_{k-l}) = \frac{\sigma(F_{k-l})}{\sigma(f_j)}$$

²This assumption implies that patterns of the form $(-f_i, -f_j)$ are likely to be below the threshold value. We therefore disregard them during the mining process.

$$= \frac{\sigma(f_1, \dots, f_{l-1}, f_{l+1}, \dots, f_k)}{\sigma(f_j)} - \frac{\sigma(F_k)}{\sigma(f_j)} \quad (1)$$

We know that

$$\{f_1, \dots, f_{l-1}, f_{l+1}, \dots, f_k\} \subset F_k$$

and

$$\frac{\sigma(F_k)}{\sigma(f_j)} = \text{MaxPI}(F_k) > t$$

Therefore from the anti-monotonic property of the support measure, it follows that

$$\frac{\sigma(f_1, f_2, \dots, f_{l-1}, f_{l+1}, \dots, f_k)}{\sigma(f_j)} > t > 0.5$$

Hence it follows from (1) that $\text{MaxPI}(F_{k-l}) < t$.

COROLLARY 3.1. Let $F = \{f_1, f_2, \dots, f_m\}$ be the set of all features in a spatial database. Then, $0 \leq P \leq m^2 - m$ where P is the size of the candidate 2-itemsets pruned because of Lemma 3.1.

When mining for positive and negative patterns, for every $f_i \in F$, we also have to consider $-f_i$. This increases the feature set size to $2m$.

Hence the number of candidate itemsets of size 2 generated from this feature set is $\binom{2m}{2}$ (because of $\max\text{PI}$ no pruning is done at the first level).

From this set we remove candidates of the form $(-f_i, -f_j)$ and $(f_i, -f_i)$. This reduce the size of the candidate set to $\binom{2m}{2} - \binom{m}{2} - m$.

Case 1 : All positive candidate item sets of size 2 are greater than the threshold t . Then by Lemma 3.1 all the candidate itemsets of size 2 with negative features would be pruned. Hence the additional number of pruning will be

$$\binom{2m}{2} - \binom{m}{2} - m - \binom{m}{2} = m^2 - m$$

Case 2 : All positive candidate item sets of size 2 are less than the threshold t .

Then all negative candidate item sets will be checked and hence there will be no additional pruning because of Lemma 3.1 .

3.1 NP_MaxPI Algorithm: Example With the help of an example we describe the details of the *Negative Pruning MaxPI* (NP_MaxPI) algorithm for mining complex spatial relationships.

No	Clique	Transaction
i	A_1, B_1, C_1	A, B, C
ii	A_2, B_2, C_2	A, B, C
iii	C_2, D_1	C, D
iv	A_3, C_3	A, C
v	A_3, B_6, D_2	A, B, D
vi	A_4, B_3, D_3	A, B, D
vii	A_5, B_5, D_4	A, B, D
viii	B_4, C_4, D_5	B, C, D
ix	A_6, D_6	A, D

Table 3: Clique Set and Transaction Set

Algorithm: NP_MaxPI

Input: Transaction table, t

Output: Confident complex patterns

```

k=1
 $F_k \leftarrow \{f_1, f_2, \dots, f_n\} \cup \{-f_1, -f_2, \dots, -f_n\}$ 
            $\cup \{f_1^+, f_2^+, \dots, f_n^+\}$ 
While  $F_k \neq \phi$ 
  k=k+1
  //generate candidate itemsets
   $C_k = \text{maxPIgen}(F_{k-1})$ 
  for each  $c \in C_k$ 
    if everyfeature in c is positive
      OR check_negative(c) = true then
        generate  $F_k$  from  $C_k$  using MaxPI
      end if
    end for
  end While

```

procedure: check_negative

Input: candidate itemset, c

Output: Boolean variable, *check*

```

check=true
for each negative feature  $-f_i \in c$ 
   $cp = c - (-f_i) + f_i$ 
  if  $cp \in F_k$  and  $f_i \notin \text{maxFeature}(cp)$  then
    check = false
  end if
end for
return check

```

Figure 2: *AlgorithmNP_MaxPI*. We generate candidate itemsets in the same way as apriori-gen(F_{k-1}), except that we use the weak anti-monotonic property instead of anti-monotonic property.

1. Given a spatial database, create a clique set such that every row represents a set of point forming a clique in the dataset. Table 3 gives an example of a clique set.
2. Create the transaction set from the clique set, in which every row represents the features of points in the corresponding row of the clique set. For example, row (i) in Table 3 gives the set of point A_1, B_1, C_1 forming a clique. Here A_1 is a spatial point with feature 'A'. Similarly B_1 is a spatial point with feature 'B' and C_1 is a spatial point with feature 'C'. Hence the transaction corresponding to this clique is $\{A,B,C\}$.
3. Generate candidate itemsets of size one with positive and negative features. For the given example, the candidate 1-itemsets are $\{A,-A,B,-B,C,-C,D,-D\}$. For simplicity (in this example), we ignore the self_co-locations such as $A+$. Since maxPI for all the 1-itemsets are 1, we do not prune any candidates at this level.
4. Generate candidate 2-itemsets from 1-itemsets, automatically pruning patterns such as $\{A, -A\}$ and patterns with all negative features such as $\{-A, -B\}$.
5. Check maxPI of the candidate 2-itemsets against the user-defined threshold, and generate frequent 2-itemsets. While checking for negative candidate itemsets use Lemma 1 for additional pruning. Tables 4 and 5 show the process of checking candidate itemsets of sizes 2 and 3 with a threshold t of 60%. The column 'checked' shows whether the candidate itemset was checked against the threshold or not. When using the MaxPI algorithm all the given candidate itemsets would have been checked. However NP_MaxPI avoids these checks because of Lemma 1. This increases the efficiency of mining complex patterns. The column *MaxFeature* de-

candidate itemset	checked	Pr	MaxPI	MaxFeature	Pruned
AB	Y	5/6,5/6	5/6	A,B	N
A-B	N				
AC	Y	3/6,3/4	3/4	C	N
A-C	Y	3/6	3/6	-	Y
AD	Y	4/6,4/6	4/6	A,D	N
A-D	N				
-AB	N				
-AC	N				
-AD	N				
BC	Y	3/6,3/4	3/4	C	N
B-C	Y	3/6	3/6	-	Y
BD	Y	4/6,4/6	4/6	B,D	N
B-D	N				
-BC	N				
-BD	N				
CD	Y	2/4,2/6	2/4	-	Y
C-D	Y	2/4	2/4	-	Y
-CD	Y	4/6	4/6	D	N

Table 4: Mining candidate 2- itemset

candidate 3-itemset	checked	pr	MaxPI	MaxFeature	Pruned
ABC	Y	2/6,2/6,2/4	2/4	-	Y
ABD	Y	3/6,3/6,3/6	3/6	-	Y
ACD	Y	0/6,0/4,0/6	0	-	Y
A-CD	Y	4/6,4/6	4/6	A,D	N
BCD	Y	1/6,1/4,1/6	1/4	-	Y

Table 5: Mining candidate 3- itemset

notes the feature for which the participation ratio of the candidate itemset was greater than the threshold. For example in Table 4, for candidate item AB, $\text{pr}(\{A,B\},A)=5/6 > t$ and $\text{pr}(\{A,B\},B)=5/6 > t$. Therefore by Lemma 1, we do not have to check $\text{maxPI}(\{-A, B\})$ and $\text{maxPI}(\{A, -B\})$ as we know that they will be less than t . If we consider AC, $\text{pr}(\{A,C\},A)=3/6 < t$ and $\text{pr}(\{A,C\},C)=3/4 > t$. Hence by Lemma 1, we do not need to check $\text{maxPI}(\{-A, C\})$. However we need to check $\text{maxPI}(\{A, -C\})$.

- Repeat steps 4 and 5 for the consequent levels until there are no more frequent itemsets.

4 Experiments, Results and Analysis

We have carried out three sets of experiments to demonstrate the scalability, applicability and the statistical correctness of NP_MaxPI.

Scalability We created a large synthetic data set in order to compare the MaxPI and the NP_MaxPI

algorithms. In particular, we wanted to test how much additional pruning was being achieved by the use of Lemma 1.

Applicability We extracted a large sample from the Sloan Digital Sky Survey Database and applied the NP_MaxPI algorithm to test the efficacy of our approach on a real data set.

Correctness We applied the NP_MaxPI algorithm on a small data set to obtain an output O which contains patterns of the form $\{A, B\}$ and $\{C, -D\}$. We then applied the Ripley's K-function test to check whether A and B are positively correlated and C and D are negatively correlated.

4.1 Performance evaluation of the algorithm

We generated synthetic datasets of sizes ranging from 170K to 1 million transactions with ten different features. These datasets were generated using the following method. With each of the twenty features (ten simple features eg. 'A' and ten self-co-locations eg. 'A+')

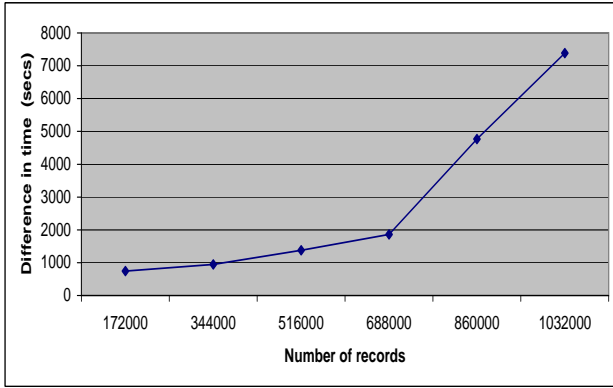


Figure 3: Comparison of efficiencies of MaxPI and *NP_MaxPI* algorithms

as the first feature, a random number of transactions were generated. Each transaction was generated with a random size from 1 to 10. Depending on the size of the transaction the consequent features were chosen randomly from the feature set.

We applied the MaxPI algorithm and *NP_MaxPI* algorithm on the datasets to generate confident complex patterns. Figure 3 shows the difference in time taken by both these methods to mine complex patterns from datasets of different sizes. It can be seen that the time difference increases rapidly with the size of the dataset, which indicates the extensive pruning of the negative patterns done by *NP_MaxPI* algorithm and hence the increase in efficiency. We confirmed that the rule sets generated by the two methods are identical by using the *diff* function.

We carried out a second experiment to compare the performance of MaxPI algorithm and *NP_MaxPI* with increasing number of features in the dataset. We generated synthetic datasets with number of features ranging from 10 to 25. Figure 4 shows the time taken by the two algorithms to mine confident complex patterns from these datasets. From the figure, it can be seen that there is a rapid increase in time with the MaxPI algorithm, while with *NP_MaxPI* the time taken is comparatively less.

4.2 Application of *NP_MaxPI* algorithm on SDSS database

4.2.1 Data Preparation The data for this experiment was obtained from the SDSS Data Release 1 online catalogue service. The online data contained information about over 150,000 celestial objects. Among the hundreds of attributes stored in the database for each object we extracted the following attributes for our ex-

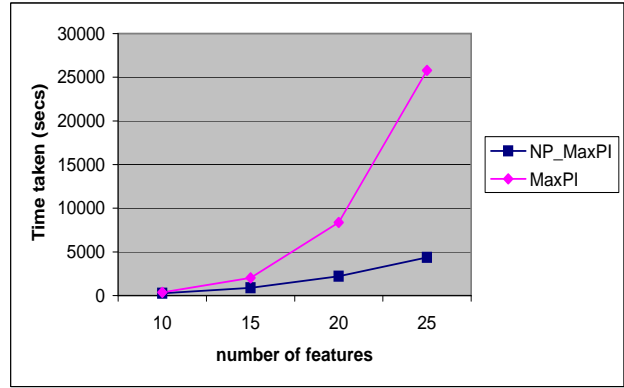


Figure 4: Comparison of efficiency with increasing number of features

periment

- The object's ID
- coordinates (as a unit vector)
- object type
- primary target flags
- red shift
- the difference between the u & r light magnitudes (used to separate galaxy types)

The distances between objects were calculated using Hubble's Law given by $D = c * z / H_0$, where c is the speed of light, z is the red shift and H_0 is the Hubble's constant and is $71 \text{ kmsec}^{-1} \text{ Mpc}^{-1}$ [4].

To ensure that the results for measuring the distance to each object would be as accurate as possible, only objects with a $z\text{Conf}$ value > 0.95 (i.e. the object's red shift is $> 95\%$ certain) and $z\text{Warning} = 0$ (i.e. there's no problem with the red shift) are used. This filtering cuts down the number of objects to around 117,000.

The 'object type' attribute classifies the objects into 25 categories. However the current online data has objects only in 17 of these categories. Among the various object types, we extracted only the galaxies since 90% of the objects were classified as 'Galaxies'. We further classified the galaxies into 'main' galaxies and Luminous Red Galaxies (LRG). The main galaxies are closer (to the Earth) and brighter than the LRG. In the SDSS database the LRG were flagged with a particular value. We classified the 'main' and LRG galaxies further into Early and Late galaxies using the UV & red light magnitudes. A $u-r$ value greater than or equal to 2.22 indicate Early galaxy and less than 2.22 indicate Late galaxies. From Hubbles Tuning Fork model [10] it

Objects	Symbol
GALAXY-LRG-EARLY	A
GALAXY-LRG-LATE	B
GALAXY-MAIN-EARLY	C
GALAXY-MAIN-LATE	D

Table 6: List of object types and their symbols

Neighbourhood Distance :1 megaparsec
min confidence : 70%
$B+ \Rightarrow -C$
$D+ \Rightarrow -C$
$A+ \Rightarrow -B - D$
$B \Rightarrow -A - C$
$C \Rightarrow -B - D$
$C+ \Rightarrow -B - D$
$A \Rightarrow -B - C - D$

Table 7: Rules from SDSS database

follows that early galaxies are elliptical in shape and late are spiral/irregular. Table 6 lists the different types of galaxies and the corresponding symbols used in this paper.

4.2.2 Rules Generated from SDSS Database

We applied the NP_MaxPI algorithm to the data set extracted from the SDSS database and some of the interesting rules generated are shown in Table 7. The entire result set could be obtained from <http://www.cs.usyd.edu.au/~chawla/sdss.html>. From Table 6 we see that features A and C are early galaxies and hence they are elliptical in shape. Features B and D are late galaxies and hence they are spiral in shape.

Among the rules in Table 7, the rules to be noted are $A+ \Rightarrow -B - D$ and $C+ \Rightarrow -B - D$. These rules show that A+ is negatively correlated with B and D but not C. Similarly C+ is negatively correlated with B and D and not A. These rules conform to the well know fact that when elliptical galaxies co-locate the spiral galaxies are excluded.

4.3 NP_MaxPI and Ripley's K-Function Ripley's K function is a method in spatial statistics which is used to characterize the spatial pattern of point data [3]. The K function is given by

$$K(t) = \lambda^{-1}E$$

where E is the number of spatial points within distance t of a randomly chosen point and λ is the density (number per unit area) of events. If A is the area of the study region and N is the observed number

of points then, $\lambda = N/A$.

Ripley's K function could be used to test complete spatial randomness i.e. to test whether the observed events are consistent with a homogeneous Poisson process. If so, $K(t) = \pi t^2$ for all t. i.e. under complete randomness $t = (K(t)/\pi)^{1/2}$ for all t. Let

$$L(t) = (K(t)/\pi)^{1/2}$$

Depending on the value of L(t), the distribution of the points could be characterized as follows:

$$L(t) = \begin{cases} < t & \text{points are negatively correlated} \\ \approx t & \text{points are randomly distributed} \\ > t & \text{points are positively correlated} \end{cases}$$

Dixon [8] has suggested the following generalization of K(t) function for multivariate spatial point process:

$$K_{ij}(t) = \lambda_j^{-1}E$$

Where E is the number of type j events within distance t of a randomly chosen type i event.

Under complete spatial randomness $K_{ij}(t) = \pi t^2$. Hence

$$L_{ij}(t) = (K_{ij}(t)/\pi)^{1/2} = t$$

This shows that values of $L_{ij}(t) > t$ indicate attraction between the i type and j type points and values $< t$ indicate repulsion.

We applied the NP_MaxPI algorithm to a very small synthetic data set with 20 points and generated a set of complex rules with *minconf* 70%. The rules are given in Table 8.

We applied Ripley's K function to the synthetic data set and calculated the L_{ij} values for different values of t where i, and j were the types of features in each of the rules in Table 8. The different values of L_{ij} are given in Table 9.

Table 9 shows that for type A and D, $L_{AD} \geq t$ which indicates that these two types of objects are positively correlated. When comparing types C and D, we find that $L_{CD} \leq t$ which shows negative correlation between the types. Similarly types C and B show negative correlation. These confirm the rules in Table 8. However it should be noted that Ripley's K-function finds relationships globally from the entire set of spatial objects whereas, our approach generates relationships locally within the neighbourhood of spatial objects.

5 Conclusion

In this paper we demonstrated the problem of generating complex patterns in spatial databases. We then presented an efficient approach, NP_MaxPI, which

Neighbourhood Distance =4
$A \rightarrow D$ conf=83.33%
$D+ \rightarrow A$ conf=100.00%
$D+ \rightarrow -C$ conf=75.00%
$C+ \rightarrow -B$ conf=100.00%

Table 8: Rules from the synthetic data set

t	L_{AD}	L_{DC}	L_{CB}
0.00	0.00	0.00	0.00
0.50	0.00	0.00	0.00
1.00	0.00	0.00	0.00
1.50	0.00	0.00	0.00
2.00	2.02	0.00	0.00
2.50	3.50	1.81	2.33
3.00	4.52	2.52	2.33
3.50	4.52	3.11	3.37
4.00	4.52	3.11	3.37

Table 9: Results from Ripley’s function

uses the maxPI measure instead of the traditional support measure for mining complex patterns. We showed that our approach effectively reduces the candidate itemsets by exploiting a complementarity property of maxPI measure which extensively prunes candidate negative patterns. We then presented and analysed various experimental results. The results of our experiments show (i) the significant performance improvement over MaxPI algorithm, (ii) the efficacy of our approach on real dataset by generating confident complex pattern in SDSS spatial database and (iii) the statistical correctness of our algorithm using Ripley’s K function.

6 Acknowledgements

Thanks to Chris Bowman for helping us create the data set. The second author is partially supported by an ARC Discovery Research Grant.

References

- [1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases VLDB*, pages 487–499. Morgan Kaufmann, 1994.
- [2] Maria-Luiza Antonie and Osmar R.Zaiane. Mining positive and negative association rules: An approach for confined rules. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD04*, 2004.
- [3] Noel A.C. Cressie. *Statistics for spatial Data*. John Wiley and Sons, 1993.
- [4] D.N.Spergel, M.Bolte, and W.Freedman. The age of the universe. *Proceedings of the National Academy of Science*, 94:6579–6584, 1997.
- [5] Yan Huang, Hui Xiong, Shashi Shekhar, and Jian Pei. Mining confident co-location rules without a support threshold. In *Proceedings of the 18th ACM Symposium on Applied Computing ACM SAC*, 2003.
- [6] Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium on Advances in Spatial Databases*, pages 47–66. Springer-Verlag, 1995.
- [7] Rob Munro, Sanjay Chawla, and Pei Sun. Complex spatial relationships. In *Proceedings of the 3rd IEEE International Conference on Data Mining, ICDM 2003*, pages 227–234. IEEE Computer Society, 2003.
- [8] P.Dixon. Ripley’s k function. department of statistics, iowa state university, 2001.
- [9] Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns:a summary of results. In *Proceedings of the 7th International Symposium on Spatial and Temporal Databases SSTD01*, 2001.
- [10] V.J.Martin and E.Saar. *Statistics of the Galaxy Distribution*. Chapman and Hall/CRC, 2002.
- [11] Xindong Wu, Chengqi Zhang, and Shichao Zhang. Mining both positive and negative association rules. In *Proceedings of 19th International Conference on Machine Learning, ICML2002*, 2002.