Mining Complex Relationships in the SDSS SkyServer Spatial Database

Bavani Arunasalam, Sanjay Chawla, Pei Sun and Robert Munro
School of Information Technologies, University of Sydney
{bavani,chawla,psun2712,rmunro}@it.usyd.edu.au

Abstract

In this paper we describe the process of mining complex relationships in spatial databases using the Maximal Participation Index (maxPI), which has a property of discovering low support and high confidence rules. Complex relationships are defined as those involving two or more of: multi-feature co-location, self-co-location, one-to many relationships, self-exclusion and multi-feature exclusion. We report our results of mining complex relationships in data extracted from the Sloan Digital Sky Survey (SDSS) database.

1. Introduction

Spatial data mining is a process of extracting interesting patterns from large spatial databases. In particular, mining co-location patterns is an important spatial data mining task. Co-location patterns represent subsets of spatial features whose instances are often located in close spatial proximity. We demonstrate the application of our spatial data mining technique to the Sloan Digital Sky Survey (SDSS) database in order to extract interesting complex spatial relationships (positive, negative and mixed) among the various celestial objects.

The Sloan Digital Sky Survey (SDSS) was established with the objective to collect and store accurate data about large number of astronomical objects in order to address a broad range of astronomical questions. The goal of the SDSS is to image all “bright” objects in 1/4 of the Northern sky, in five different wavelengths of light. The SDSS project is now underway having begun its standard operations in April 2000, and is planned to last for five years [3].

A spatial relationship is one between features in an Euclidean space, defined in terms of the co-locational trends over that space. An example is to determine the confidence of the ‘where there’s smoke there’s fire’ with respect to a set of coordinates, each representing the feature, smoke or fire. The task here would be to determine whether fire occurs in the neighbourhood of smoke more than is randomly likely.

Neighbourhoods are defined in terms of cliques (also known as neighbour-sets). A clique is defined as any set of items such that all items in that set co-locate. For example, in Figure 1 and Table 1, the co-locational pattern \{A, C\} occurs 3 times, v, vii and x. In spatial data, two features are typically said to co-locate if they are positioned within a distance $d$ of one another. As has been assumed in Figure 1, $d$ is usually a constant, but it may also be defined as varying locally within the space or with respect to a given feature.

The mining of association rules is a well-developed field. However most of the methods developed for mining confident cliques as transactions, fail to capture many spatial phenomena of interest, as most mining techniques have been optimized for ‘marketbasket’ data. Spatial data is fundamentally different from market-basket data, both in its basic nature and distributional tendencies.

One factor unique to spatial data is that the number of transactions a single item may participate in is potentially unbounded, while in a market-basket this is limited to one (obviously, two people may purchase the same toothpaste product, but not the same exact tube). Self-co-location is also more likely in spatial data. The upper limit in a marketbasket is multiple purchases of only one product, which is less likely than an equivalent spatial situation of an area of
monoculture forest. Similarly, there may be direct relationships between features that don’t co-locate, as between animals displaying territorial behaviour, making such relationships intrinsically more interesting in spatial data.

Perhaps the most fundamental difference between spatial and other data in a transactional representation is the notion of significance. A co-location is considered significant if it occurs more than is randomly likely. In transactions representing market-baskets, the transactions, by definition, represent the complete space of the data (there are no empty baskets). In such cases, the significance of the data may be represented by frequency of the features in relation to the number of transactions. In spatial data, however, the random likelihood of a co-location depends on the volume of the space from which it was taken.

2. Maximal Participation Ratio

In this section we will briefly describe the notion of Maximal Participation Index (maxPI) as described in [4] where more details can be found.

2.1. Participation ratio

Given a co-location pattern \( L \) and a feature \( f \in L \), the participation ratio of \( f \), \( \text{pr}(L, f) \), 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 \( \text{pr}(\{A, B, C\}, C) = \frac{2}{6} \).

2.2. Maximal participation index

Given a co-location pattern \( L \), the maximal participation index of \( L \), \( \text{maxPI}(L) \) can be defined as the maximal participation ratio of all the features in \( L \), that is:

\[
\text{maxPI}(L) = \max_{f \in L} \{ \text{pr}(L, f) \}.
\]

For example, in figure 1, \( \text{maxPI}(\{A,B,C\}) = \max(\text{pr}(\{A,B,C\}, A), \text{pr}(\{A,B,C\}, B), \text{pr}(\{A,B,C\}, C)) = \max(\frac{2}{5}, \frac{2}{3}, \frac{2}{6}) = \frac{2}{3} \).

A high maximal participation index indicates that at least one spatial feature strongly implies the pattern. By using maxPI, rules with low frequency but high confidence can be found, which would be pruned by a support threshold.

2.3. Mining rules with low support and high confidence using maxPI

As discussed above the maxPI 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 \( A \rightarrow B \), which is given as
\[
\text{conf}(A \rightarrow B) = \frac{\text{support of } \{A,B\}}{\text{support of } \{A\}},
\]
could be high.

The \( \text{maxPI} \) directly addresses the problem. \( \text{maxPI} \) is given as
\[
\text{maxPI of } \{A,B\} = \max(\text{pr}(\{A,B\},A),\text{pr}(\{A,B\},B))
\]
This shows that a high confidence for the rule \( A \rightarrow B \) will lead to high \( \text{maxPI} \), which will prevent rules with low support and high confidence being pruned.

2.4. The weak monotonic property of \( \text{maxPI} \)

Maximal participation index is not monotonic with respect to the pattern containment relations. For example, in figure 1, \( (\text{maxPI}(\{A,C\}) = 3/5 < (\text{maxPI}(\{A,B,C\})=2/3 \). Interestingly, as pointed out in [4] the maximal participation index does have the following weak monotonic property:

If \( R \) is a \( k \)-colocation pattern, then there exists at most one \((k-1)\) subpatterns \( R' \) of \( R \) such that \( \text{maxPI}(R') < \text{maxPI}(R) \).

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

3. Notations used

**Absence:** The absence of an item is represented by negation, for example \( \neg A \). (Note: this is not the equivalent of the set-theory, \( \neg A \), meaning the presence of any item other than \( A \)).

**Self-co-location:** Multiple instances of a feature (multiple items) are represented by a ‘+’ following the feature, for example \( A+ \).

4. Algorithm for Mining complex relationships using maximal participation index

Following are the steps involved in mining complex relationships in the SDSS database:
1. For every instance (record) in the data set get all its neighbours within a specified distance\(^1\). The neighbours of each instance form a clique and hence a transaction. Therefore the number of transactions in the data set is equal to the number of instances.
2. To every clique we add features representing the absence of items and the presence of multiple items.
3. Apply the \( \text{maxPI} \) algorithm to the transactions, automatically pruning trivial/nonsensical co-locations such as \( A \rightarrow \neg A \) and \( A+ \rightarrow A \). Return a set frequent of co-locations with a specified minimum threshold.
4. Generate rules from the form the frequent cliques with a specified minimum confidence.

5. Data Preparation and Experiments

The data for this experiment was obtained from the SDSS Data Release 1, online catalogue service. This data is now available in a Microsoft SQL Server format [3]. The online data contained information about over 150000 celestial objects. The objects are classified into 25 categories. However the current online data has objects in 17 of these categories only. Among these objects, we extracted only the galaxies because 90% of the objects were classified as ‘Galaxy’. 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. From Hubbles Tuning Fork model [5] it follows that early galaxies are elliptical in shape and late are spiral / irregular. Table 2 lists the different types of galaxies and the corresponding symbol used in this paper.

6. Results

We applied the \( \text{maxPI} \) algorithm to the data set extracted from the SDSS database and some of the interesting rules generated are shown in Table 3. The entire result set could be obtained from [http://www.cs.usyd.edu.au/~chawla/SDSS.html](http://www.cs.usyd.edu.au/~chawla/SDSS.html). From Table 2 we see that Feature 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. The results of our experiment conform to the well know fact that when elliptical galaxies co-locate the spiral galaxies are excluded.

\(^1\) This is not strictly a clique. However, while being easier to program, it does not seem to have much effect on the results.
Table 2: List of object types and their symbols

<table>
<thead>
<tr>
<th>Objects</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALAXY_LRG_EARLY</td>
<td>A</td>
</tr>
<tr>
<td>GALAXY_LRG_LATE</td>
<td>B</td>
</tr>
<tr>
<td>GALAXY_MAIN_EARLY</td>
<td>C</td>
</tr>
<tr>
<td>GALAXY_MAIN_LATE</td>
<td>D</td>
</tr>
<tr>
<td>HOT_STD</td>
<td>E</td>
</tr>
<tr>
<td>QSO</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 3: The results

<table>
<thead>
<tr>
<th>Neighbourhood distance $d$: 1 megaparsec</th>
</tr>
</thead>
<tbody>
<tr>
<td>min confidence: 70%</td>
</tr>
</tbody>
</table>

A+ $\Rightarrow$ -B-D-F conf = 71.9254  
C+ $\Rightarrow$ -B-D-F conf = 90.13

7. Conclusion

We have reported preliminary results about the application of spatial data mining techniques to extract potentially useful patterns from SDSS database. Our results show that our methods are capable of extracting well-known relationships from the database.

8. Acknowledgements

Thanks to Chris Bowman for helping us create the data set.

9. References


