Modeling Spatiotemporal Data With Contrast Frequent Patterns

Dawei Wang
Outline

• Introduction to Patterns
  • Frequent Pattern and Contrast Frequent Pattern
  • Pros and Cons
  • Application

• Modeling Spatial Temporal Data
  • Motivation and Challenges
  • Methodology
    ➢ Visualization with reasoning
    ➢ Prediction through classification
  • Modeling Practice
Frequent Patterns:

A rule defined by its **frequency** in the data transactions

<table>
<thead>
<tr>
<th>Transactions</th>
<th>$T_1$ : ${AR = high, POP = low, IC = low}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_2$ : ${AR = high, POP = low, IC = high}$</td>
</tr>
<tr>
<td></td>
<td>$T_3$ : ${AR = high, POP = low, IC = medium}$</td>
</tr>
<tr>
<td></td>
<td>$T_4$ : ${AR = medium, POP = low, IC = medium}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patterns</th>
<th>$X_1$ : ${AR = high, POP = low}$</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_2$ : ${AR = high, IC = high}$</td>
<td>$\text{sup}(X_1) = \frac{3}{4} = 75%(T_1, T_2, T_3)$</td>
</tr>
<tr>
<td></td>
<td>$X_3$ : ${AR = high}$,</td>
<td>$\text{sup}(X_2) = \frac{1}{4} = 25%(T_2)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\text{sup}(X_3) = \frac{3}{4} = 75%(T_1, T_2, T_3)$</td>
</tr>
</tbody>
</table>

Table 1: Examples of transactions, patterns and patterns’ supports. Pattern $X_3$ is not a closed pattern because $X_1$, its immediate superset, has exactly the same support. $X_1$ is a closed frequent pattern if we set the minimum support threshold $\alpha = 70\%$.

- Based on Binary Data!!
Frequent Patterns Mining

• NP-Hard Problem

• Apriori Algorithm

• FP_Growth Algorithm

Apriori Principle:

In frequent items
Pruned supersets
Contrast Frequent Pattern

• Designed for classification related task\(^1\)
• Defined by two parameters: **support** and **growth ratio**


Lots of names: Emerging Pattern/Contrast Set/Subgroup Mining\(^2\)
Facts about Contrast Pattern

- **G**: Growth Ratio
- **S**: Support in the partition
- **N**: Number of Instances in the partition

$$\text{False Negative (FN)} = N(1-S);$$  \hspace{1cm}  \text{Precision} = \frac{G}{G+1};
$$\text{True Positive (TP)} = N* S;$$  \hspace{1cm}  \text{Recall} = S;
$$\text{False Positive (FP)} = \frac{(N*S)}{G};$$  \hspace{1cm}  \text{F1 score} = \frac{2GS}{GS+S+G};

In the non-perfect real word data sets, high support usually leads to low growth ratio.
Pros and Cons

Cons:

• NP-Hard for mining
  • Inapplicable to data with high dimensions

• Statistical meaningless
  • 100 random patterns, expecting 10 with 90% support

• Too many patterns
  • How to summarize the results is one of the remaining hotspots in pattern mining research

• Lack of flexibility for representing the regularities within the data
  • Binary format
  • Arbitrary thresholds

Pros:

• Intuitive data representation
• Great ability for abstraction
Application

• Sequential pattern (e.g., time series)
• Spatial Co-location pattern
• Recommendation algorithms (association rules)
Modeling Spatiotemporal Data

1. Motivation and Challenges
   ➢ What is spatiotemporal data and why is it interesting?

2. Methodology
   ➢ Visualization with reasoning
   ➢ Prediction through classification

3. Modeling Practice:
   ➢ Crime Data
   ➢ Climate Science Data
Spatiotemporal Data

- Data instances with labels showing where and when they were collected.

- Both the spatial and temporal information of observations contribute to the analytical purposes.
Why is it interesting?

Ex.1: Crime Analysis

- Crime incidence data: where and when the crime happened.

- Accurately identified and clearly visualized crime hotspots can significantly benefit crime analysis and police practices.
**Why is it interesting?**

Ex2: Climate Science Data

<table>
<thead>
<tr>
<th>Meteorological Variables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Z300</td>
<td>300hPa Geopotential Height</td>
</tr>
<tr>
<td>Z500</td>
<td>500hPa Geopotential Height</td>
</tr>
<tr>
<td>Z1000</td>
<td>1000hPa Geopotential Height</td>
</tr>
<tr>
<td>U300</td>
<td>300hPa Zonal Wind</td>
</tr>
<tr>
<td>V300</td>
<td>300hPa Meridional Wind</td>
</tr>
<tr>
<td>U850</td>
<td>850hPa Zonal Wind</td>
</tr>
<tr>
<td>V850</td>
<td>850hPa Meridional Wind</td>
</tr>
<tr>
<td>T850</td>
<td>850hPa Temperature</td>
</tr>
<tr>
<td>PW</td>
<td>Precipitable water</td>
</tr>
</tbody>
</table>

- Multiple variables, each of which having multiple levels in longitude, latitude, and altitude, and multiple time phases;
- **Prediction** of natural disasters like flood and tornado can greatly benefit society;
- “Big Data”, but relatively few studies from the perspective of data science*.

With the above variables sampled during a certain **temporal period** at these **locations**, can we **forecast** the extreme rainfalls happened in the ECA area?

Model I

Visualization with Reasoning
Challenges for Mapping Hotspot

- How to avoid using hard thresholds?

- Crime is understood to be related to a variety of socio-economic and crime opportunity variables, how can we integrated these related information into hotspot mapping?

- Are all the hotspots the same? Can we cluster them?
Methodology (1)
Visualization with Reasoning

• Understanding the spatial distribution of spatial incidents based on their related variables

• Modeling the spatiotemporal data using transaction-based geospatial dataset and design a data representation (GDPattern) for summarize the patterns crime related variables.

• To avoid hard thresholds for hotspot, define hotspots based on an optimization approach
  ➢ Develop the algorithm, Hotspot Optimization Tool (HOT), to utilize GDPattern for optimizing the visualization of spatial incidents' hotspots

• Cluster the hotspot based on similarity measure of its GDPatterns.
From Contrast Frequent Pattern to Geospatial Discriminative Pattern (**GDPattern**)

- Define transactions based on the spatial coordinates

<table>
<thead>
<tr>
<th>Transactions</th>
<th>Patterns</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$: ${AR = high, POP = low, IC = low}$</td>
<td>$X_1$: ${AR = high, POP = low}$</td>
<td>$sup(X_1) = \frac{3}{4} = 75% (T_1, T_2, T_3)$</td>
</tr>
<tr>
<td>$T_2$: ${AR = high, POP = low, IC = high}$</td>
<td>$X_2$: ${AR = high, IC = high}$</td>
<td>$sup(X_2) = \frac{1}{4} = 25% (T_2)$</td>
</tr>
<tr>
<td>$T_3$: ${AR = high, POP = low, IC = medium}$</td>
<td>$X_3$: ${AR = high}$</td>
<td>$sup(X_3) = \frac{3}{4} = 75% (T_1, T_2, T_3)$</td>
</tr>
</tbody>
</table>

$$\delta = \frac{sup_{D_h}(X)}{sup_{D_n}(X)}$$, Growth Ratio of Pattern $X$
HOT: Optimizing through Reasoning

• Optimizing user-specified hotspot boundaries using GDPattern:
  1. Identify areas with a relatively high crime density ($D_{h0}$);
  2. Generate candidate locations (lines 8-12): locations located in $D_{h0}$ and adjacent to some cell in $D_h$.
  3. Test the hypothesis for candidate cells (line 14): a candidate cell is inside the footprints of GDPatterns mined from $D_h$.
  4. If the hypothesis is true, the boundaries of the hotspot are modified by changing the current cell into a hotspot cell (from $D_{h0}$ to $D_h$) (line 15);
  5. Iterate until all hypothesis tests are fault (line 3 and line 19).

Algorithm 1. The hotspot Optimization Tool.

**Data:**
- $h$: a hotspot threshold
- $h'$: a hotspot candidate threshold
- $\rho$: a support threshold of closed frequent pattern
- $\delta$: a growth ratio threshold

**Result:**
- $D_h$: a new set of hotspots
- $G$: a set of GDPatterns
- $\psi$: GDPattern footprints

```plaintext
1 count = 1;
2 Generate $D_h$, $D_{h'}$, and $D_h$;
3 while count $\neq 0$ do
4     count = 0;
5     $\mu = \emptyset$;
6     $G = $ Mine GDPatterns using $D_h$, $\rho$ and $\delta$;
7     $\psi = $ footprints$(G)$;
8     for cell $c \in D_{h'}$ do
9         if $c$ adjacent to some cell in $D_h$ and $c \in D_h'$ then
10            $\mu = \mu \cup c$;
11        end
12     end
13     for cell $c \in \mu$ do
14         if $c \in \psi$ then
15             $D_h = D_h \cup c$;
16             count++;
17         end
18     end
19 end
```
Clustering through Similarity Measure*

Case 1: \[ s(X_i, Y_i) = \frac{2 \times \log P(X_i \vee Z_1 \vee Z_2 \cdots \vee Z_k \vee Y_i)}{\log P(X_i) + \log P(Y_i)} \]

Case 2: \[ s(\cdot, Y_i) = \sum_{k=1}^{n} P_X(Z_k) s(Z_k, Y_i) \]

Case 3: \[ s(\cdot, \cdot) = \sum_{l=1}^{n} \sum_{k=1}^{n} P_X(Z_l) P_Y(Z_k) s(Z_l, Z_k) \]

where \( P() \) is the probability calculated using the known distribution of the values \( i_{th} \) variable in \( D \) and \( Z_1, Z_2, \ldots, Z_k \) is the ordinal interval delimited by \( X_i \) and \( Y_i \). For example, in the left figure the ordinal interval between the first variable \( A \) in patterns \( X \) and \( Y \) is \( Z_1 = 2 \).

Study the Spatial Distribution of Real World Crime Data using \textit{GDPattern} and \textit{HOT}

• A National Institute of Justice funded project for crime study.

• Publication:
Understanding the spatial distribution of crime based on its related variables

1. Introduce a spatial data mining concept, Geospatial Discriminative Patterns (GDPatterns), to study the relationship between target crime hotspots and their underlying related variables.

2. Introduce a model, Hotspot Optimization Tool (HOT), to identify crime hotspots through their related variables.

3. Visualize the locations of those clusters in a rational way to assist domain scientists in further analysis, using the footprints of GDPatterns.

Understanding the spatial distribution of crime based on its related variables

1. Results of HOT

- High residential burglary (RB) rates are associated with high population density (POP) only in areas with few foreclosures (FC), commercial burglaries (CB), motor-vehicle larcenies (MV), street robberies (SR), and very low arrest rates (AR). These areas also have high residential density (HU) and are close to universities or colleges (DC). Such locations are shown in the footprint map.

2. Visualization of Crime Hotspot Clusters
Model II

Prediction Through Classification
The main challenge for studying spatiotemporal data is to **model the** spatial / temporal variable-variable and variable-label **relationships**.

The fact of multiple variables and multiple spatial/temporal levels for each variable usually results in an **extreme high dimensional** feature space and **complex interactions and correlations** (e.g., spatial autocorrelation)

“Rare events” in temporal scale (e.g. flood may only happened once per year) results in an extreme imbalanced data set.
1. Mining location based contrast temporal patterns of individual variables. 
   **Summarize on the temporal dimension**

2. Growing individual variable patterns to clusters. 
   **Summarize on the spatial dimension**

3. Pattern Based Classifier For Prediction 
   **Summarize the interactions among variables**

* Dawei Wang, Wei Ding, A Hierarchical Framework for Learning Climate Science Data and Forecasting Extreme Weather Events, submitted to ICDM 2015
Step1: Data Representation: Contrast Pattern

For every variable on every location, generate one new feature (summarize on the temporal scale).

<table>
<thead>
<tr>
<th></th>
<th>V¹</th>
<th>V²</th>
<th>V³</th>
<th>V⁴</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>I₁</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>P</td>
</tr>
<tr>
<td>I₂</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>P</td>
</tr>
<tr>
<td>I₃</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>I₄</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>P</td>
</tr>
<tr>
<td>I₅</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>N</td>
</tr>
</tbody>
</table>

Pattern X: \{<V₁,1>, <V₂,0 to 4>, <V₃,1 or 2>, <V₄,3>,\}

P-support=2/3=0.67
N-support=1/2=0.5
Growth Ratio = 0.67/0.5=1.34

<table>
<thead>
<tr>
<th>Feature of Pattern X</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>I₁</td>
<td>1</td>
</tr>
<tr>
<td>I₂</td>
<td>1</td>
</tr>
<tr>
<td>I₃</td>
<td>0</td>
</tr>
<tr>
<td>I₄</td>
<td>0</td>
</tr>
<tr>
<td>I₅</td>
<td>1</td>
</tr>
</tbody>
</table>
Classify an instance by querying its spatial cluster patterns.

- if we found 5 spatial clusters ($C^1, C^2, C^3, C^4, C^5$) using SCOT and an testing instance only containing the first and the second clusters ($\{< C^1, 1 >, < C^2, 1 >, < C^3, 0 >, < C^4, 0 >, < C^5, 0 >\}$), we would calculate the growth ratio of this pattern in the training data set and make predictions based on this information.
Forecasting the Extreme Rainfall Event in the Eastern Central Andes (ECA)*

• Definition of *Extreme Rainfalls*
  • In a 48 hours period, if the total rainfall volume within the ECA area is above the 99th percentile of historical records.

• Prediction Variables
  • 16 variables, including pressure, temperature, humidity, and wind at different vertical levels.
  • All variables are sampled at 1281 spatial locations, with 3-hours intervals.
  • Given 8 time stamps, we are going to have $1281 \times 16 \times 8 = 163,968$ features.

• Prediction Challenge
  • If there is an extreme rainfall events in the coming 48 hours, can we use the last 24 hours data (8 time stamps) to accurately predict it?
Output of SCOT:

- From individual location based time series pattern to spatial cluster
The patterns found in **fig. a** and **fig. b** conform with domain study.*
Presenter:

Dawei Wang

- PhD Candidate in Computer Science
  - Department of Computer Science, UMass Boston
  - September 2011 – present
  - Research Areas: Data Mining and Machine Learning