Rafał Zarajczyk

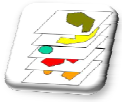
Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data

Short introduction
Input data description

- Set of spatial data divided into layers
- Each layer corresponds to the certain *attribute*
- The value of the attribute for a certain point is *true* (1) if the point lies within region (or cluster – group of points)

Short introduction
Input data description - example

- Each layer may contain data such as:
  - Railway stations (points)
  - Crime incidents (points)
  - Parks (area)
  - Urban area (area)
  - Schools (points)
  - Police stations (points)

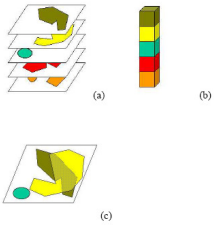


Short introduction
What do we want to get?

- Find patterns of concentration data on one layer in relation to data on other layers
- Amount of data is large, so we have to consider computational expensiveness
- We don't have any prior information and domain knowledge – we can't give any hypothesis about patterns

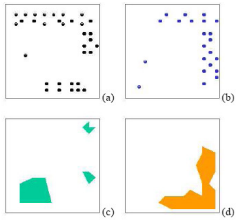
Introduction (maybe not so short?)
Two approaches

- Vertical-view approach
  - We 'cut out' one vertical bar from out layers
  - Depicted on picture (b)
- Horizontal-view approach
  - We overlay all layers onto each other
  - Depicted on picture (c)



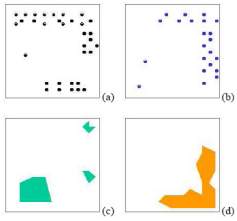
Vertical-view approach
Introduction

- Let's assume we have following data available:
  a) Railway stations locations
  b) Crime incidents locations
  c) Park areas
  d) Urban areas



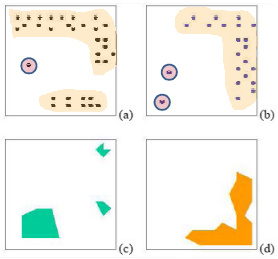
Vertical-view approach
Introduction

- Data are
  - On different layers
  - Both point based and area based



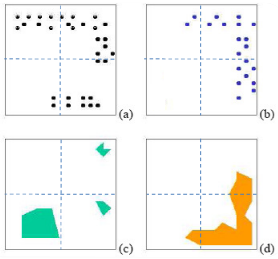
Vertical-view approach
Clusters and noise points

- Identify *clusters* and *noise points*
  - *Clusters* – groups of points
  - *Noise points* – single points, without neighbors
- How? Using any cluster analysis method
- Noise points are ignored



Vertical-view approach
Adding grid

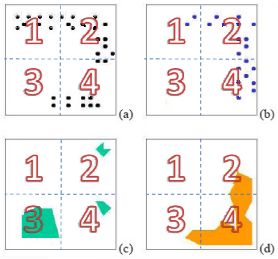
- Add grid with collectively exhaustive and mutually exclusive cells
- Number of cells is a parameter
- Let's use 2×2 grid



Vertical-view approach
Computing relational table

	Layer (a)	Layer (b)	Layer (c)	Layer (d)
Loc 1	1	1	0	0
Loc 2	1	1	1	1
Loc 3	1	0	1	1
Loc 4	1	1	1	1

1 – there are parts of region/cluster
0 – there are no parts of region/cluster



Vertical-view approach
Mining multivariate associations

	Layer (a)	Layer (b)	Layer (c)	Layer (d)
Loc 1	1	1	0	0
Loc 2	1	1	1	1
Loc 3	1	0	1	1
Loc 4	1	1	1	1

- It is a transaction table!
- So it is easy to mine associations using any method from this table

Digression
Some definitions

- Notation: $X \Rightarrow Y\ (c\%)$
  - That mean: *c% of data that satisfy X also satisfy Y*
  - *c* is called *confidence*
- Definitions
  - *confidence* is an estimate for: $Pr[X \cap Y] / Pr[X]$
    - Conditional probability of Y given X
  - *support* is an estimate for: $Pr[X \cap Y]$
    - Ratio of transactions that satisfy both X and Y to the number of all transactions

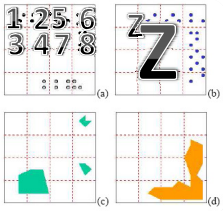
Vertical-view approach
Mining multivariate associations

	Layer (a)	Layer (b)	Layer (c)	Layer (d)
Loc 1	*1*	*1*	*0*	*0*
Loc 2	***1***	***1***	***1***	***1***
Loc 3	1	0	1	1
Loc 4	***1***	***1***	***1***	***1***

1 – there are points
0 – there are no points

- One of mined rules:
  $layer(a) \wedge layer(b) \Rightarrow layer(d)\ (66.7\%)$
- What does it mean?
  - 66.7% locations, that are near-by railway stations and have crime incidents, fall under urban areas
  - Support of this is 50%

Vertical-view approach
What if we change grid?



	Layer(a)	Layer(b)	Layer(c)	Layer(d)
Loc 11	1	0	0	0
Loc 12	1	1	0	0
Loc 13	0	0	0	0
Loc 14	0	0	0	0
Loc 21	1	1	0	0
Loc 22	1	1	1	0
Loc 23	0	0	0	0
Loc 24	1	1	0	1
Loc 31	0	0	1	0
Loc 32	0	0	1	0
Loc 33	0	0	1	0
Loc 34	1	0	1	1
Loc 41	0	0	0	0
Loc 42	1	1	1	1
Loc 43	1	0	0	1
Loc 44	1	1	0	1

Vertical-view approach
What if we change grid?

- Our previous rule
- Confidence is now 50%
- But support decreases to 18.8%

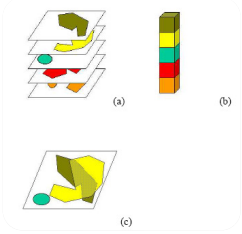
	Layer(a)	Layer(b)	Layer(c)	Layer(d)
Loc 11	1	0	0	0
Loc 12	*1*	*1*	*0*	*0*
Loc 13	0	0	0	0
Loc 14	0	0	0	0
Loc 21	*1*	*1*	*0*	*0*
Loc 22	*1*	*1*	*1*	*0*
Loc 23	0	0	0	0
Loc 24	***1***	***1***	***0***	***1***
Loc 31	0	0	1	0
Loc 32	0	0	1	0
Loc 33	0	0	1	0
Loc 34	1	0	1	1
Loc 41	0	0	0	0
Loc 42	***1***	***1***	***1***	***1***
Loc 43	1	0	0	1
Loc 44	***1***	***1***	***0***	***1***

Vertical-view approach
Summary

- Advantage
  - Easy to apply transactional association-rule mining techniques
- Disadvantage
  - Highly dependent on the granularity that is difficult to determine

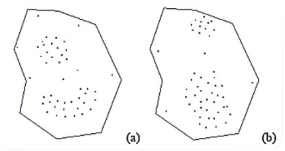
Layered data model once again
Two approaches

- Vertical-view approach
  - We 'cut out' one vertical bar from out layers
  - Depicted on picture (b)
- Horizontal-view approach
  - We overlay all layers onto each other
  - Depicted on picture (c)

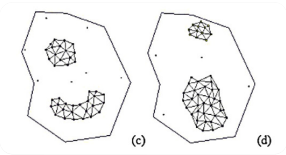


Horizontal-view approach
Introduction

- Now our input data look like this
  a) Dataset I (46 points)
  b) Dataset II (50 points)
- Area of region: 6940.14



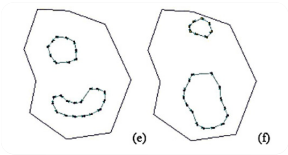
Horizontal-view approach
Cluster detection



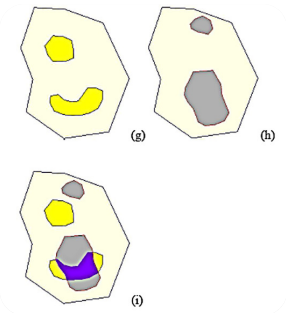
- Detect clusters and noise points using boundary-based cluster detection algorithm (Estivill-Castro and Lee)
- Again, noise points are ignored (Lee)

Horizontal-view approach
Cluster boundary extraction

- Apply the cluster boundary extraction process
- Then we polygonize clusters and the area inside them



Horizontal-view approach
Overlaying clusters



- Finally we put all clusters on one layer
- We can visually see that areas intersect
- But how to define association?

Digression no 2
More definitions

- Let X be a set of layers
- *cluster_areas(X)*
  - If X is a single point-data layer : set of polygonized clusters of X
  - Else: the total area of regions that result of the intersection of *cluster_areas*(X_i), for all X_i in X
- *Clusters with Ratio R of P (CwR(P))*
  - Clusters detected by a clustering algorithm whose normalized sizes (number of points / total number of points) are greater or equal than R

Digression no 2
More definitions

- Clustered Spatial Association Rule (CSAR): expression in the form of

$$X \Rightarrow Y\ (CC\%), \quad for\ X \cap Y = 0$$

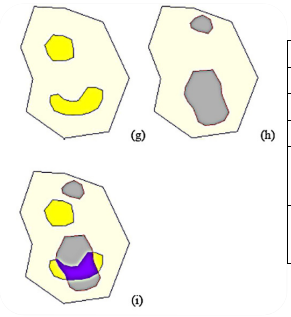
- This means: $CC\%$ of areas of clusters of X intersect with areas of clusters of Y

Digression no 2
Even more definitions

$$X \Rightarrow Y\ (CC\%), \quad for\ X \cap Y = 0$$

- Clustered Support – CS: ratio of area that satisfy both X and Y to the area of study region S
  - $CS = (\ clusters\_area(X) \cap clusters\_area(Y)\ ) / area(S)$
- Clustered Confidence – CC: conditional probability of areas of CwR of Y given areas of CwR of X
  - $CC = clusters\_areas(X \cup Y) / clusters\_areas(X)$

Horizontal-view approach Calculating rules



	clusters_area	CS(%)	CC(%)
S	6940.14	100.0	N/A
Dataset I	992.04	14.29	N/A
Dataset II	1312.21	18.91	N/A
Dataset I ⇒ Dataset II	401.46	5.78	40.47
Dataset II ⇒ Dataset I	401.46	5.78	30.59

Horizontal-view approach Calculating rules

- Around 40% of locations belonging to clusters in Dataset I also belongs to clusters in Dataset II
  - Around 40% of incidents illustrated in Dataset I happens near incidents from Dataset II
- Vice-versa similar

	clusters_area	CS(%)	CC(%)
S	6940.14	100.0	N/A
Dataset I	992.04	14.29	N/A
Dataset II	1312.21	18.91	N/A
Dataset I ⇒ Dataset II	401.46	5.78	40.47
Dataset II ⇒ Dataset I	401.46	5.78	30.59

Horizontal-view approach Summary

- Advantages
  - Autonomous – better suited for mining massive databases than the vertical-view approach
  - Does not necessitate domain knowledge
- Disadvantages
  - ???

Real data example Introduction

- Crime activity on the south east Queensland region
- 217 suburbs around Brisbane
- Crime data provided by Queensland Police Services are too complex and extremely huge
  - It is difficult even for domain experts to detect valuable patterns

Real data example Input data

- Queensland Police Service provides data:
  1. Offences against person
     - Homicide, assault, sexual offence, robbery, extortion, kidnapping, others
  2. Offences against property
     - Breaking and entering, arson, other property damage, motor vehicle theft, stealing, fraud, others
  3. Other offences
     - Drug offences, prostitution, liquor, gaming offences, trespassing and vagrancy, good order offences, traffic and related offences, miscellaneous offences

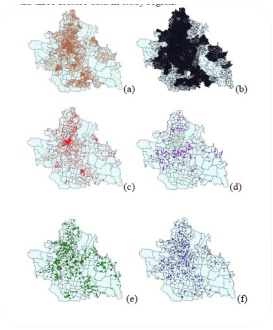
Real data example Even more input data

- Parks
- Railway stations
- Schools
- Other features

- To our purposes we will use 3 main crime categories and 3 feature data

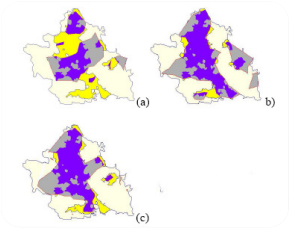
Real data example
Input data selection

a) Offences against person – 9 618 cases
b) Offences against property – 113 618 cases
c) Other offences - 2 124 cases
d) Reserves
e) Parks (including caravan parks)
f) Schools



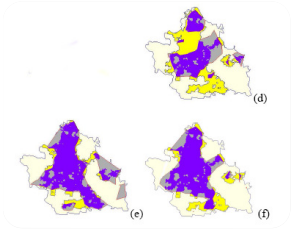
Real data example
Clustering and polygonization

- Overlaps between *cluster_areas(offences against the person)* and
  a) cluster_areas(reserves)
  b) cluster_areas(parks)
  c) cluster_areas(schools)



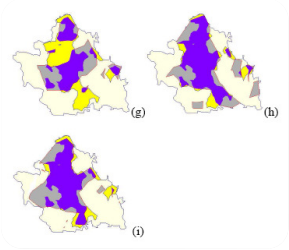
Real data example
Clustering and polygonization

- Overlaps between *cluster_areas(offences against the property)* and
  d) cluster_areas(reserves)
  e) cluster_areas(parks)
  f) cluster_areas(schools)



Real data example
Clustering and polygonization

- Overlaps between *cluster_areas(other offences)* and
  g) cluster_areas(reserves)
  h) cluster_areas(parks)
  i) cluster_areas(schools)



Real data example
Quantitatively described data

	CS(%)	CC(%)
Offences Against the person ⇒ Reserves	15.40	44.93
Reserves ⇒ Offences Against the person	15.40	50.99
Offences Against the person ⇒ Parks	29.23	85.29
Parks ⇒ Offences Against the person	29.23	57.33
Offences Against the person ⇒ Schools	26.56	77.50
Schools ⇒ Offences Against the person	26.56	59.85
Offences Against the property ⇒ Reserves	20.83	47.44
Reserves ⇒ Offences Against the property	20.83	68.99
Offences Against the property ⇒ Parks	36.25	82.56
Parks ⇒ Offences Against the property	36.25	71.10
Offences Against the property ⇒ Schools	33.42	76.11
Schools ⇒ Offences Against the property	33.42	75.31
Other offences ⇒ Reserves	17.81	50.47
Reserves ⇒ Other offences	17.81	58.97
Other offences ⇒ Parks	29.90	84.74
Parks ⇒ Other offences	29.90	58.64
Other offences ⇒ Schools	28.35	80.36
Schools ⇒Other offences	28.35	63.89

Real data example
What can we read from that?

- The amount of CSARs is really big
- Let's filter data and choose only these, where CS minimum is 30%, and CC minimum is 75%
  – Offences against property ⇒ Parks (36.25% CS, 82.56% CC)
  – Offences against property ⇒Schools (33.42% CS, 76.11% CC)
  – Schools ⇒ Offences against property (33.42% CS, 76.31% CC)

Real data example
Final conclusions

- Most offences against property are taking place around parks and schools
- Locations of school will probably cause offences against property
- If you live near some school in Queensland – beware!

To finish with…
Summary

Vertical-view approach

1. Find spatial clusters for point-data layers
2. Segment all layers with the finite number of regular cells
3. Construct m×n relational table with the binary values
4. Apply association-rule mining to the table

Horizontal-view approach

1. Find *CwR(P)* for point-data layers P in X and Y
2. Extract clusters boundaries of each *CwR* for point-data layers in X and Y
3. Compute the value of the areas of *CwR* for point-data layers and the areas of area-data layers
4. Overlay X and Y
5. Apply association-rule mining to detect CSARs

The end

- Thank you for your attention