Term Project: Crime forecasting

Assigned Date: Thursday, April 21, 2011

Educational Goal

Apply data mining techniques to real-world crime data.

Team

Team 1	Team 2
Ward, Max W Yu, Chung-Hsien Das, Priyank	Zhu, Bingchun Phan, Dung Huy Le, Hie

Phase IV: Empirical study for crime prediction

(200 points)

Due: 5:30 PM Tuesday, May 17, 2011

Requirements

- **Dataset**: A dense version two years (2006, 2007) aggregated dataset is available. The dataset is in the same format as previous assignments. The grid size is 48*40.
  Dataset description: Each CSV file is a matrix containing 48 rows and 40 columns. Those 48*40 grids cover the geographical region of city Boston. Each entry of this matrix reflects the number of a certain crime incidents.
  The coordinate file coord.CSV has the latitude and longitude of those 48*40 grids. Each line in the dataset is correspondent to a coordination pair. Suppose A is the 48*40 grid matrix, then the coordination of A(i,j) can be found at(j-1)*48+i line in coord.CSV file (i and j start from 1).
- **Team 1 classification approach**

Team 1 will focus on finding crime hotspots (have more than 1 crime) in January 2007 (1 for hotspot and 0 for coldspot). A very desirable task is predicating the tendency of crimes to be increased or decreased in an area, creating additional features based on neighborhood count and crime change rate, and applying feature selection. The team is encouraged to explore your ideas in the last phase of the term project.

There are three steps:

1. Training and test data generation → 2. Classification → 3. Results visualization

Step 1. Training and Test Data Generation.

The team is required to propose a new method or modify an existing method on training set and test set generation. Data generation should involve the locality information based on neighboring grids. Compare the proposed method with the previous approaches (constrained prediction and unconstrained prediction) you have done in previous phases, using the new data set provided in this phase.

Step2. Classification process

Use the same set of classifiers used to in phase 3 to calculate the precision, recall, F1 measure and accuracy for the prediction of hotspots. Please exclude any grid cell that does not include any crime data when calculating the prediction results.

Step3. Results visualization

1) show the precision, recall, F1, accuracy compared with previous approaches.
2) Demonstrate the results using Google Earth API.
3) A simple example of ArcGIS Silverlight visualization for the prediction results.
4) Discuss the advantages and the disadvantages the methods the team has explored in this phase.

- ## Team 2 regression approach

Team 2 will focus on predicating the exact number of Burglary crimes in January 2007.

There are three steps:

1. Training and test data generation-> 2. Regression -> 3. Results visualization

Step1.Training and test data generation

Based on the lessons learned and experiences gained in previous phases, the team should explore their own ideas on training and test data generation. The team is encouraged to involve **other crimes information**.

Step 2. Regression process

Use the same regression method in phase 2 to calculate the mean square error (MAE) for the prediction of hotspots. Please exclude any grid cell that does not include any crime data when calculating the prediction results.

Part 3. Results visualization

1) compare the mean square error (MAE) with the previous approaches on this new grid sizes.
2) Demonstrate the results using ArcGIS Silverlight visualization for the prediction results.
3) Discuss the advantages and the disadvantages of the new approaches based on your observation.

Submission Requirements

1. Write an experiment report to discuss your experimental results, including detailed parameter settings and **figures** for the visualization results. Submit the paper copy of the report, and source

code and **KML** of the scripts with the cover page in class. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.

2. Submit the softcopy of the report and scripts through your UMass Online account athttp://boston.umassonline.net/index.cfm.
3. Zip all the files. One submission per team. Save the file as sdm_ teamNumber. For example, Team 1 should name their file as *sdm_team1.zip.*
4. No hard copies or soft copies results in 0 points.