Term Project: Crime forecasting

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 III: Empirical study for crime prediction

(200 points)

Assigned Date: Tuesday, April 5, 2011

Due: 5:30 PM Tuesday, April 19, 2011

Requirements

- **Dataset**: The datasets contains 6 crime types and each crime is presented in monthly base in CSV (Comma-Separated Values) file format. **The data can only be used for this course project. Please collect it from the TA after signing a non-spreading agreement** (send an email to yangmu@cs.umb.edu to make an appointment).
  Note that you should not open the CSV files directly using MS Excel Spreadsheet because the data would be crashed if you do so. Read the files using a programming language, for example, Java or Matlab.

  Dataset description:
  Training data is a collection of 6 crime incidents (Streetrobbery, ResidentialBurglary, mv_larceny, Foreclousure, CommercialBurglary, Arrest) from January to December in 2006. Please note that each training data file contains crime data for one particular month. Test data is for ResidentialBurglary crime in January 2007.
  Each CSV file is a matrix containing 24 rows and 20 columns. Those 24*20 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 24*20 grids. One line is a coordination pair. Suppose A is the 24*20 grid matrix, then the coordination of A(i,j) can be found at(j-1)*24+i line in coord.CSV file (i and j start from 1).

- ## Team 1 and Team 2

  Statistics of the Crime Data: Apply various statistical analyses to the data using the following approaches but not limited to:

  - Scatter plots comparing residential burglary counts to the other attribute counts to evaluate correlations
  - Analysis of variance between the counts and residential burglary
  - Line graphs of crime counts during the year to get a sense of the cycles of different crimes (look for highs and lows of different crime types during different times of the month and year)
  - Any other statistical analysis to help you understand the data

- ## Team1 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).

  There are three steps: 1. Training and test data generation → 2. Classification → 3. Results visualization

  Step 1. Training and Test Data Generation.

  In phase II, we used the ***t*** month based prediction, where ***t*** is ranging from 1 to 10. Then we learned the best ***t*** which achieves the best F1 measure. In phase III, we adopt the model of leave one month out (LOMO) to predict the results. You should use the best ***t*** learned in phase II. Note that in phase III, The test month will be among January 2006, February 2006, …, December 2006, and January 2007. For example, if the best ***t*** learned in phase II is two-month based (***t***=2), then the starting month of the test month in phase III should be March 2006 because you will use January and February for training which is ***t***+1. LOMO means each single month may be extracted as the test data and all the previous month(s) may be used to build training data. In summary, test data starts from ***t+1*** and ends at January 2007 (totally ***12-t+1*** test data; actually you already had the results of January 2007 in phase II). The way to build training data and test data is the same as previous.

  Step 2. Classification Process

  Use the same classifiers you used before in phase III to calculate the precision, recall, F1 measure and accuracy for the prediction. **However, in phase III, you should remove the grid cell that has never had a crime before (most likely in ocean), then excluding those grid cells when calculating the prediction results.**

  Step3. Results Visualization

1) For each classifier you use, we have ***12-t+1*** results (including January 2007) when applying LOMO method. Totally, we have results from four classification algorithms. We can draw a figure which shows the relation of prediction results versus months (***t***). In the figure there are four curves corresponding to four algorithms. Four values are required to shown in the figures: precision, recall, F1, accuracy.
2) Google earth visualization for ***12-t+1*** results. Use different buttons to represent different months.
3) Discuss the crime tendency of each month and prediction results. Discuss your conclusion and observation.

- ## 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

In phase II, we used the ***t*** month based prediction, where ***t*** is ranging from 1 to 10. Then we learned the best ***t*** which achieves the best F1 measure. In phase III, we adopt the model of leave one month out (LOMO) to predict the results. You should use the best ***t*** learned in phase II. Note that in phase III, The test month will be among January 2006, February 2006, …, December 2006, and January 2007. For example, if the best ***t*** learned in phase II is two-month based (***t***=2), then the starting month of the test month in phase III should be March 2006 because you will use January and February for training which is ***t***+1. LOMO means each single month may be extracted as the test data and all the previous month(s) may be used to build training data. In summary, test data starts from ***t+1*** and ends at January 2007 (totally ***12-t+1*** test data; actually you already had the results of January 2007 in phase II). The way to build training data and test data is the same as previous.

Step 2. Regression Process

Use the same regression method in phase II to calculate the mean square error (MAE) for the prediction of hotspots. **However, in phase III, you should remove the grid cell that has never had a crime before (most likely in ocean), then excluding those grid cells when calculating the prediction results.**

Step 3. Results Visualization

1) For each regression algorithm, we have ***12-t+1*** results when applying LOMO method. Totally, we have four algorithms. We can draw a figure which shows the relation of prediction results versus months (***t***). In the figure there are four curves corresponding to four algorithms. One figure is required to illustrate MAE versus months.
2) Google earth visualization for ***12-t+1*** results. Use different buttons to represent different months.

3) Discuss the crime tendency of each month and prediction results. Discuss the prediction results of different regression algorithms 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 at http://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.