Term Project: Crime forecasting

Assigned Date: Thursday, March 10, 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 II: Empirical study for crime prediction

(200 points)

Due: 5:30 PM Thursday, March 31, 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).

- ## 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

Step1 Training and Test Data Generation.

In phase 1, we generated the training and test data according to one month based prediction. In this phase, we use the ***t*** month based prediction, where ***t*** is ranging from 1 to 10. Simply concatenating the ***t*** months' features will form a much longer vector to represent one training sample. Accordingly, the total number of training samples will decrease. We need to build 10 groups training and test datasets by different ***t***.

Step2. Classification process

In phase 1, we tested 1NN classifier. In this phase, we should choose 3 other different classification methods including SVM, J48 (Decision tree), and Artificial Neural Network in Weka. Calculate the precision, recall, F1 measure and accuracy for the prediction of hotspots.

Step3.Results visualization

1) For each algorithm, we have 10 results (each dataset will have one result). Totally, we have 4 algorithms. We can draw a figure which shows the relation of prediction results versus months (***t***). In the figure there are 4 curves corresponding 4 algorithms. 4 figures are required: precision, recall, F1, accuracy.
2) Discuss the influence for the prediction results when pre-known month changes. Discuss the prediction results of different classification algorithms based on your 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 1, we generated the training and test data according to one month based prediction. In this phase, we use the ***t*** month based prediction, where ***t*** is ranging from 1 to 10. Simply concatenating the ***t*** months' features will form a much longer vector to represent one training sample. Accordingly, the total number of training samples will decrease. We need to build 10 groups training and test datasets by different ***t***.

Step 2. Regression process

In phase 1, we tested simple linear regression classifier. In this phase, we should choose 3 other different regression methods of your choice in Weka. Calculate the mean square error (MAE).

Part 3. Results visualization

3) For each regression algorithm, we have 10 results (each dataset will have one result). Totally, we have 4 algorithms. We can draw a figure which shows the relation of prediction results

versus months (***t***). In the figure there are 4 curves corresponding 4 algorithms. 1 figure illustrating MAE versus months is required.

4) Discuss the influence for the prediction results when pre-known month changes. 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 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.