On Mining Anomalous Patterns in Road Traffic Streams

Linsey Xiaolin Pang[1,4] Sanjay Chawla[1] Wei Liu[2] Yu Zheng[3]

[1] School of Information Technologies, University of Sydney, Australia
[2] Dept. of Computer Science and Software Engineering, University of Melbourne, Australia
[3] Web Search and Mining Group, Microsoft Research Asia, Beijing, China
[4] NICTA, Sydney, Australia
qlinsey@it.usyd.edu.au, sanjay.chawla@sydney.edu.au,
wei.liu@unimelb.edu.au, yuzheng@microsoft.com

December 17, 2011

Introduction

- Background
  - Huge volumes of spatio-temporal data are available.
  - Detection of abnormal traffic patterns is helpful.
- Procedures–how to detect the top anomalous regions
  - Statistical Significance Computation: For any given region, the hypothesis test is applied to calculate the score of a region, typically LRT(likelihood ratio test statistical)
  - Searching: An efficient approach is required to detect spatial-temporal outliers. Naive approach is very time-consuming.

Proposed Approach

- Approaches
  - A generic framework for spatio-temporal outlier detection based on existing LRT work is proposed.
  - Persistent and emerging outlier detection models are defined in our work.
  - We prove that the pruning strategy of LRT is suitable in persistent and emerging scenarios

Model Definitions

- PSTO Model (Persistent Spatio-Temporal Outlier Model):

$$D(R) = \begin{cases} \dfrac{\Pi_{r_i \in R} L(\theta_r|X_R) \Pi_{r_i \in \bar{R}} L(\theta_{\bar{r}}|X_{\bar{r}})}{\Pi_{r_i \in G} L(\theta_G|X_G)} & \text{for } \theta_r \geq \theta_{\bar{r}}\,, \\ 1 & \text{otherwise.} \end{cases}$$

- ESTO Model (Emerging Spatio-Temporal Outlier Model):

$$D(R) = \begin{cases} \dfrac{Max_{\theta_{\bar{r}} \leq \theta_{t_{min}} \leq \ldots \leq \theta_T} \Pi_{r_i \in R} L(\theta_r^t|X_r^t) \Pi_{r_i \in \bar{R}} L(\theta_{\bar{r}}^t|X_{\bar{r}}^t)}{\Pi_{r_i \in G} L(\theta_G^t|X_G^t)} & \text{for } \theta_{\bar{r}} \leq \theta_{t_{min}} \\ 1 & \text{otherwise.} \end{cases}$$

Upper-bounding and Pruning Mechanism

- Lemma: Let region $R = R_{t1} \cup R_{t2}$ for non-overlapping time interval $t1$ and $t2$, we have:

$$L(\theta_R|X_R) \leq L({\theta'}_{R_{t1}}|X_{R_{t1}}) \times L({\theta'}_{R_{t2}}|X_{R_{t2}}) \tag{1}$$

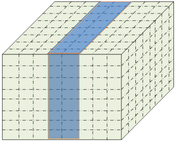
, where $\theta_R = \theta_{R_{t1}} \cup \theta_{R_{t2}}$ and $X_R = X_{R_{t1}} \cup X_{R_{t2}}$

- Lemma: Let region $R = R1 \cup R2$ for non-overlapping spatial region R1 and R2,we have:

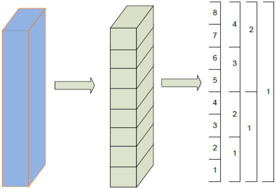
$$L(\theta_{R1}, \theta_{R2}|X_{R1}, X_{R2}) \leq L({\theta'}_{R1_{t1}}, {\theta'}_{R1_{t2}}|X_{R1_{t1}}, X_{R1_{t2}}) \times L({\theta'}_{R2_{t1}}, {\theta'}_{R2_{t2}}|X \tag{2}$$

,where R, $R1$, $R2$ are composed of (t1,t2) time steps respectively. Here we just use two time steps to illustrate. It is applicable to any t time steps.

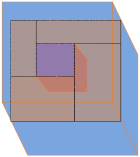
Upper-bounding and Pruning Mechanism

- Upper-bounding
- 

(a) R



(b) R



(c) $\bar{R}$

Figure: Precomputation of any given spatial-temporal region R and tiling of $\bar{R}$.

Computational Complexity

- In brute-force approach, there are totally $O(n^6)$ regions to be searched in space-time dimension and the overall cost is $O(cn^6)$.
- Our approach reduce the cost by pre-compute two likelihood data set: $O(n^4)$,$O(n^3)$.

Experiment Results

- Synthetic Data
  - The results are investigate from three aspects: (a) average pruning rate; (b) accuracy; (c) average running time.
  - Scenario I :The null hypothesis holds.
  - Scenario II :The null hypothesis holds. The data in a random selected cuboid area with size of $5 \times 4 \times 3$ is generated with different parameter setting.
  - Scenario III: The alternative hypothesis holds (subtle outlier).
  - Scenario IV: The alternative hypothesis holds (extreme outlier). The data of a randomly selected cuboid area with size of $5 \times 4 \times 3$ was generated by different success rate.

Experiment Results

- Synthetic Data

$Test$	$Pruning(\%)$	$Accuracy(\%)$
$4 \times 4 \times 4$	100	no false alarm
$8 \times 8 \times 8$	100	no false alarm
$16 \times 16 \times 16$	99.9	0.1 false alarm

Table: Average Pruning Rate in Scenario I

$Test$	$Pruning(\%)$	$Accuracy(\%)$
$4 \times 4 \times 4$	100	no false alarm
$8 \times 8 \times 8$	99.99	0.01 false alarm
$16 \times 16 \times 16$	100	no false alarm

Table: Average Pruning Rate and Accuracy in Scenario II

Experiment Results

- Synthetic Data

Test	16/16/16	32/16/16	64/16/16	32/32/32	128/16/16
ppsto (%)	95.27	97.35	97.64	97.47	96.74
pesto (%)	98.37	98.46	98.69	99.11	99.23

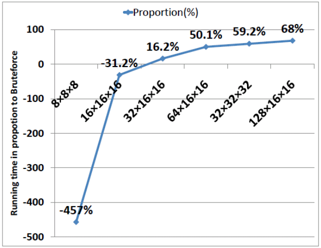
Table: Average Pruning Rate in Scenario III

Table: Average Pruning Rate in Scenario IV

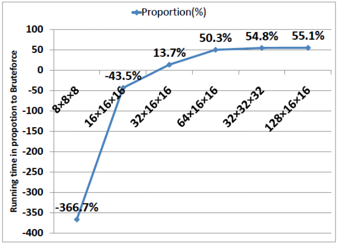
Test	16/16/16	32/16/16	64/16/16	32/32/32	128/16/16
ppsto (%)	79.27	97.51	97.77	97.22	96.68
pesto (%)	95.57	97.40	96.78	94.70	95.23

Experiment Results

- Synthetic Data



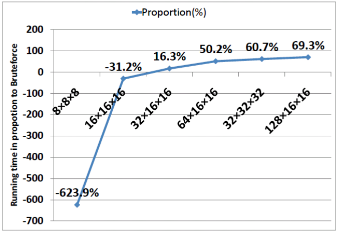
(a) Scenario III psto



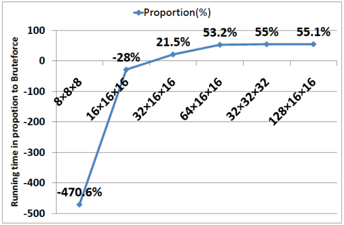
(b) Scenario III esto

Experiment Results

- Synthetic Data



(c) Scenario IV psto

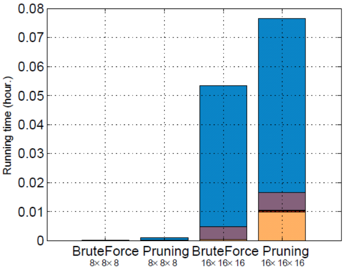


(d) Scenario IV esto

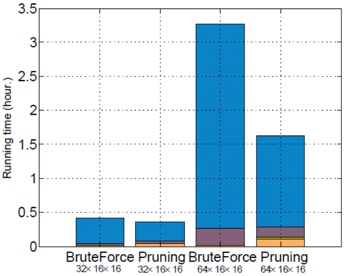
Figure: The proportion of running time of pruning vs. brute-force approach.

Experiment Results

- Synthetic Data



(a) Split cost of ESTO with smaller dataset

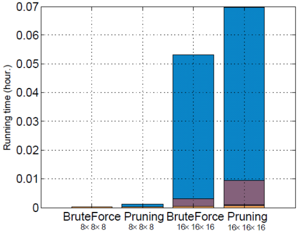


(b) Split cost of ESTO with larger dataset

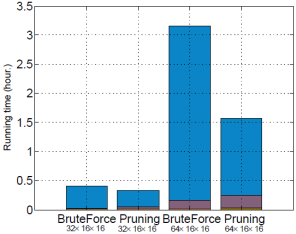
$\bar{R}$ Computation R Computation $\bar{R}$ Precomputation R Precomputation Rest Computation

Experiment Results

- Synthetic Data



(d) Split cost of PSTO with smaller dataset



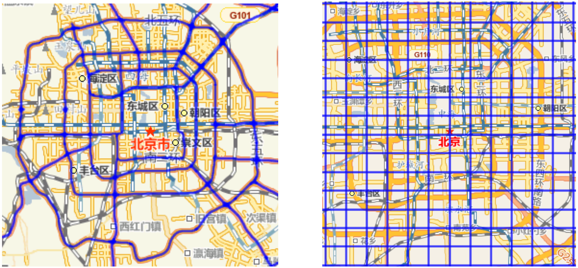
(e) Split cost of PSTO with larger dataset

$\bar{R}$ Computation, R Computation, $\bar{R}$ Precomputation, R Precomputation, Rest Computation

Figure: The running time of comparable parts of brute-force vs. pruning approach in scenario III.

Real Data

- Beijing Map



(a) Road Network (b) Grid Map

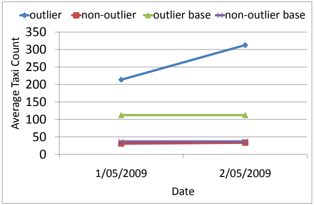
Figure: An example of the traffic network of Beijing. Based on the longitude and latitude, the entire city is partitioned into a grid map. Subfigure(a) is partitioned into subfigure(b).

Real Data

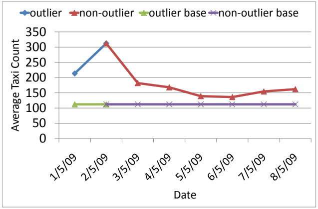
- Two cases of emerging outliers detected on a real GPS trajectory dataset generated by $33,000$ taxis in Beijing from $01/03/2009$ to $31/05/2009$.
- **Case I:** The data spans 16 days starting from $01/05/2009$ to $16/05/2009$ within 9:00:00 am to 10:00:00 am every day.
- **Case II:** The data spans 8 days starting from $14/03/2009$ to $21/03/2009$ within 3:15:00 pm to 4:30:00 pm every day.

Experiment Results

- Real Data



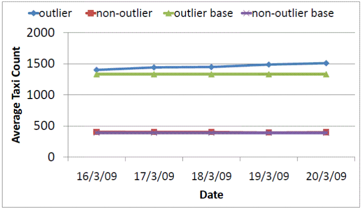
(a) The average taxi counts within outlier regions vs. non-outlier regions from 01/05/2009 to 02/05/2009



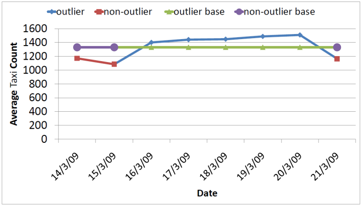
(b) The average taxi counts within outlier regions from 01/05/2009 to 08/05/2009

Experiment Results

- Real Data



(c) The average taxi counts within outlier regions vs. non-outlier regions from 16/03/2009 to 20/03/2009



(d) The average taxi counts within outlier regions from 14/03/2009 to 21/03/2009

Figure: Comparison of outlying and non-outlying regions in $8 \times 8 \times 8$ grid.

Experiment Results

- Real Data

(a) The region highlighted with blue borders on the map is the outlier region of Case I. The icon shows the exact location of Happy Valley.

(b) The region highlighted with blue borders is the outlier of Case II. It is the city express road of Beijing. (i.e. Tonghuihe North Road)

Figure: Outlier Locations from our two case studies on Beijing Map

Wu, M., Song, X., Jermaine, C., Ranka, S., Gums, J.: A LRT Framework for Fast Spatial Anomlay Detection, in KDD'09: : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 887–896.

S. S. Wilks. The large sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics, 9:60-62, 1938.

Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K.: Detection of emerging space-time clusters, in KDD'05, pp. 218–227 (2005)

Neill, D.B., Moore, A.W.: Detection of emerging space-time clusters:prior work and new directions, Technical report, Carnegie Mellon University, 2004.

R. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng., 14(5):1003–1016, 2002

- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD 1996, pages 226–231,Sept. 1996.
- W. Wang, J. Yang, and R. R. Muntz. Sting: A statistical information grid approach to spatial data mining. In VLDB 1997, pp. 186–195, 1997.
- J.Zhao,C. T. Lu,Y. F. Kou, and J. Yang: Detecting Region Outliers in Meteorological Data. GIS03, November 7-8, 2003, New Orleans, Louisiana, USA.
- M. Kulldorff. A spatial scan statistic. Comm. in stat.: Theory and Methods, 26(6):1481-1496, 1997.
- M. Kulldorff. Spatial scan statistics: models, calculations, and applications. In J. Glaz and N. Balakrishnan, editors, Scan Statistics and Applications, pp.303–322, Birkhauser, 1999.

M. Kulldorff, W. Athas, E. Feuer, B. Miller, and C. Key. Evaluating cluster alarms: a space-time scan statistic and cluster alarms in los alamos. American Journal of Public Health, 88:1377–1380, 1998.

M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. Statistics in Medicine, 14:799–810, 1995.

J. H. Friedman and N. I. Fisher. Bump hunting in high-dimensional data. Stat. and Comp., 9(2):123–143,April 1999.

V. Chandola, A. Banerjee and V. Kumar , Anomaly Detection: A survey,. Acm Computing Surveys 41,3, pp. 1–58,2009.

D. B. Neill and A. W. Moore. A fast multi-resolution method for detection of significant spatial disease clusters. In NIPS 2003, pp. 651–658, 2003.

D. B. Neill and A. W. Moore. Rapid detection of significant spatial clusters. In SIGKDD 2004, pp. 256265, 2004.

D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu. Spatial scan statistics: approximations and performance study. In SIGKDD 2006,pp. 24-33, 2006.

D. Agarwal, J. M. Phillips, and S. Venkatasubramanian. The hunting of the bump: On maximizing statistical discrepancy. In SODA 2006, pp. 1137–1146, 2006.

T. Tango, K. Takahashi, and K. Kohriyama. A SpaceTime Scan Statistic for Detecting Emerging Outbreaks. In Journal of the International Biometrics Society,67(1):106–115,2010

- L. Huang, M. Kulldorff, and D. Gregorio. A Spatial Scan Statistic for Survival Data. In Journal of the International Biometrics Society,63(1):109–118,2007
- I. Jung, M. Kulldorff and AC. Klassen: A spatial scan statistic for ordinal data. Stat Med, 26: 1594–1607, 2007.
- I. Jung, M. Kulldorff and OJ. Richard: A spatial scan statistic for multinomial data. Stat Med. Aug 15;18:1910-1918, 2010.
- L. Huang, R. Tiwari, M. Kulldorff, J. Zou, and E. Feuer : Weighted normal spatial scan statistic for heterogenous population data, in Journal of the American Statistical Association, 2009.

W. Liu, Y. Zheng, S. Chawla, J. Yuan and X. Xie : Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams, in KDD '11 17th SIGKDD conference on Knowledge Discovery and Data Mining, pp. 1010–1018,2011.

J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, Y. Huang : T-Drive: Driving Directions Based on Taxi Trajectories, in Proceedings of the 18th ACM SIGSPATIAL Conference on Advances in Geographical Information Systems, pp. 99–108,2010.