CORRECTING AND COMPLEMENTING FREEWAY TRAFFIC ACCIDENT DATA USING MAHALANOBIS DISTANCE BASED OUTLIER DETECTION

Bin Sun, Wei Cheng, Guohua Bai, Prashant Goswami

A huge amount of traffic data is archived which can be used in data mining especially supervised learning. However, it is not being fully used due to lack of accurate accident information (labels). In this study, we improve a Mahalanobis distance based algorithm to be able to handle differential data to estimate flow fluctuations and detect accidents and use it to support correcting and complementing accident information. The outlier detection algorithm provides accurate suggestions for accident occurring time, duration and direction. We also develop a system with interactive user interface to realize this procedure. There are three contributions for data handling. Firstly, we propose to use multi-metric traffic data instead of single metric for traffic outlier detection. Secondly, we present a practical method to organise traffic data and to evaluate the organisation for Mahalanobis distance. Thirdly, we describe a general method to modify Mahalanobis distance algorithms to be updatable.

Keywords: *Accident Data, Data Labelling, Differential Distance, Mahalanobis Distance, Outlier Detection, Traffic Data, Updatable Algorithm*

1 Introduction

Full-text will be available (open access):
http://dx.doi.org/10.17559/TV-20150616163905

First author's web:
http://ABOUT.DMML.NU

The code is available on GitHub:
https://github.com/SunnyBingoMe/sun2017correcting-github

7 References

[1] Guo, J.; Huang, W.; Williams, B. M. Real time traffic flow outlier detection using short-term traffic conditional variance prediction. // Transportation Research Part C: Emerging Technologies, 50(2015), pp. 160–172.

[2] Xuesong, W.; Qiang, G.; Shanshan, L.; Rongfei, C. Design and Implementation of School Hospital Information Analysis and Mining System. // Applied Science, Materials Science and Information Technologies in Industry, 513(2014), pp. 498–501.

[3] Prasad, N.; Kumar, P.; Naidu, M. M. An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree. // Fourth International Conference on Intelligent Systems, Modelling and Simulation / Bangkok, 2013, pp. 56–60.

[4] Bhatt, A. S. Comparative Analysis of Attribute Selection Measures Used for Attribute Selection in Decision Tree Induction. // International Conference on Radar, Communication and Computing / SKP Engineering College Tiruvannamalai, 2012, pp. 230–234.

[5] Using Supervised Learning and Comparing General and ANTI-HIV Drug Databases Using Chemoinformatics. // Pattern Recognition and Machine Intelligence, Proceedings / Taneja Shweta; Raheja Shipra; Kaur Savneet. Berlin: Springer-Verlag, 2009, pp. 177–183.

[6] Jiawei, H.; Kamber, M. Data mining: concepts and techniques, 3rd ed. Singapore: Elsevier, 2012.

[7] Zhao, M.; Zhan, C.; Wu, Z.; Tang, P. Semi-Supervised Image Classification Based on Local and Global Regression. // IEEE Signal Processing Letters, 22, 10(2015), pp. 1666–1670.

[8] Honig, A.; Howard, A.; Eskin, E.; Stolfo, S. Adaptive Model Generation: An Architecture for Deployment of Data Mining-based Intrusion Detection Systems. // Applications of data mining in computer security, 8720(2002), pp. 153–194.

[9] Thomas, T.; Van Berkum, E. C. Detection of incidents and events in urban networks. // IET Intelligent Transport Systems, 3, 2(2009), pp. 198–205.

[10] Anomaly Detection. // Introduction to Data Mining / Pang-Ning Tan; Michael Steinbach; Vipin Kumar. Boston: Pearson Addison Wesley, 2006, pp. 651–665.

[11] Hall, F. L.; Shi, Y.; Atala, G. On-line testing of the McMaster incident detection algorithm under recurrent congestion. // Transportation Research Record, 1394(1993), pp. 1–7.

[12] Yuan, F.; Cheu, R. L. Incident detection using support vector machines. // Transportation Research Part C-Emerging Technologies, 11, 3(2003), pp. 309–328.

[13] Ghosh-Dastidar, S.; Adeli, H. Wavelet-Clustering-Neural Network Model for Freeway Incident Detection. // Computer-Aided Civil and Infrastructure Engineering, 18, 5(2003), pp. 325–338.

[14] Ahmed, F.; Hawas, Y. E. A Threshold-Based Real-Time Incident Detection System for Urban Traffic Networks. // Transport Research Arena, 48(2012), pp. 1713–1722.

[15] Gonzalez, H.; Han, J.; Ouyang, Y.; Seith, S. Multidimensional Data Mining of Traffic Anomalies on Large-Scale Road Networks. // Transportation Research Record, 2215, 1(2011), pp. 75–84.

[16] Lin, W. H.; Daganzo, C. F. A Simple Detection Scheme for Delay-Inducing Freeway Incidents. // Transportation Research Part A-Policy and Practice, 31, 2(1997), pp. 141–155.

[17] Stephanedes, Y. J.; Chassiakos, A. P. Freeway incident detection through filtering. // Transportation Research Part C: Emerging Technologies, 1, 3(1993), pp. 219–233.

[18] Wang, Y.; Zhang, C. Alternative Route Strategy for Emergency Traffic Management Based on Its: A Case Study of Xi'an Ming City Wall. // Tehnicki Vjesnik-Technical Gazette, 20, 2(2013), pp. 359–364.

[19] Yeon, J.; Hernandez, S.; Elefteriadou, L. Differences in Freeway Capacity by Day of the Week, Time of Day, and Segment Type. // Journal of Transportation Engineering-ASCE, 135, 7(2009), pp. 416–426.

[20] Li, J.; Zhang, H. M. Fundamental Diagram of Traffic Flow - New Identification Scheme and Further Evidence from Empirical Data. // Transportation Research Record, 2260(2011), pp. 50–59.

[21] Kamarianakis, Y.; Gao, H. O.; Prastacos, P. Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions. // Transportation Research Part C-Emerging Technologies, 18, 5(2010), pp. 821–840.

[22] Vlahogianni, E. I.; Karlaftis, M. G.; Golias, J. C. Statistical methods for detecting nonlinearity and non-stationarity in univariate short-term time-series of traffic volume. // Transportation Research Part C-Emerging Technologies, 14, 5(2006), pp. 351–367.

[23] Multivariate forecasting methods. // Econometric Forecasting / Robert M. Kunst. Vienna: Institute for Advanced Studies, 2012, pp. 31–40.

[24] De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. // Chemometrics and Intelligent Laboratory Systems, 50, 1(2000), pp. 1–18.

[25] Cho, S.; Hong, H.; Ha, B.-C. A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction. // Expert Systems with Applications, 37, 4(2010), pp. 3482–3488.

[26] Krishnaswamy, J.; Bawa, K. S.; Ganeshaiah, K. N.; Kiran, M. C. Quantifying and mapping biodiversity and ecosystem services: Utility of a multi-season NDVI based Mahalanobis distance surrogate. // Remote Sensing of Environment, 113, 4(2009), pp. 857–867.

[27] Zhang, Y.; Huang, D.; Ji, M.; Xie, F. Image segmentation using PSO and PCM with Mahalanobis distance. // Expert Systems with Applications, 38, 7(2011), pp. 9036–9040.

[28] Cunderlik, J. M.; Burn, D. H. Switching the pooling similarity distances: Mahalanobis for Euclidean. // Water Resources Research, 42, 3(2006), p. 3409.

[29] Farber, O.; Kadmon, R. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. // Ecological Modelling, 160, 1–2(2003), pp. 115–130.

[30] Roess, R. P.; Prassas, E. S.; McShane, W. R. Traffic Engineering, 4th ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[31] Hypothesis Testing. // An Introduction to Mathematical Statistics and Its Applications / Richard J. Larsen; Morris L. Marx. 5th ed. Boston: Pearson, 2012.

[32] Euclidean distance. // The Cambridge dictionary of statistics / Brian Everitt. Cambridge: Cambridge University Press, 2002, p. 134.

[33] Qi, Y.; Smith, B. L. Identifying nearest neighbors in a large-scale incident data archive. // Transportation Research Record, 1879, 1(2004), pp. 89–98.

[34] Zhijun, H.; Chuangwen, X. A Kind of Algorithms for Euclidean Distance-Based Outlier Mining and its

Application to Expressway Toll Fraud Detection. // International Asia Conference on Informatics in Control, Automation and Robotics / Los Alamitos, 2009, pp. 414–417.
[35] Maurer, C. R.; Qi, R. S.; Raghavan, V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. // IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, 2(2003), pp. 265–270.
[36] Laurikkala, J.; Juhola, M.; Kentala, E.; Lavrac, N.; Miksch, S.; Kavsek, B. Informal identification of outliers in medical data. // Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology / Berlin, 2000, pp. 20–24.
[37] Filzmoser, P.; Garrett, R. G.; Reimann, C. Multivariate outlier detection in exploration geochemistry. // Computers & Geosciences, 31, 5(2005), pp. 579–587.
[38] Rousseeuw, P. J.; Van Zomeren, B. C. Unmasking multivariate outliers and leverage points. // Journal of the American Statistical Association, 85, 411(1990), pp. 633–639.
[39] Rousseeuw, P. J.; Driessen, K. V. A fast algorithm for the minimum covariance determinant estimator. // Technometrics, 41, 3(1999), pp. 212–223.
[40] Krzywicki, A.; Wobcke, W. Exploiting Concept Clumping for Efficient Incremental E-Mail Categorization. // Advanced Data Mining and Applications Pt II, 6441(2010), pp. 244–258.
[41] Sun, C. C.; Chilukuri, V. Dynamic Incident Progression Curve for Classifying Secondary Traffic Crashes. // Journal of Transportation Engineering, 136, 12(2010), pp. 1153–1158.
[42] R Core Team. The R Project for Statistical Computing. R. 1993. URL: http://www.r-project.org/. (17.05.2015.).
[43] Song, Y. E.; Stein, C. M.; Morris, N. J. strum: an R package for structural modeling of latent variables for general pedigrees. // Bmc Genetics, 16(2015), p. 35.
[44] Palarea-Albaladejo, J.; Antoni Martin-Fernandez, J. zCompositions - R Package for multivariate imputation of left-censored data under a compositional approach. // Chemometrics and Intelligent Laboratory Systems, 143(2015), pp. 85–96.
[45] Ahlin, C.; Stupica, D.; Strle, F.; Lusa, L. medplot: A Web Application for Dynamic Summary and Analysis of Longitudinal Medical Data Based on R. // Plos One, 10, 4(2015), p. 121760.
[46] Janos, S.; Martinović, G. Web based distant monitoring and control for greenhouse systems using the Sun SPOT modules. // 7th International Symposium on Intelligent Systems and Informatics 2009, pp. 165–169.
[47] Alder, J. R.; Hostetler, S. W. Web based visualization of large climate data sets. // Environmental Modelling & Software, 68(2015), pp. 175–180.
[48] Wang, R.; Zhong, D.; Zhang, Y.; Yu, J.; Li, M. A Multidimensional Information Model for Managing Construction Information. // Journal of Industrial and Management Optimization, 11, 4(2015), pp. 1285–1300.
[49] Winston Chang; Joe Cheng; J. J. Allaire; Yihui Xie; Jonathan McPherson. Shiny: Web Application Framework for R. Feb-2015. URL: http://cran.r-project.org/web/packages/shiny/index.html. (17.05.2015.).

Authors' addresses

Bin Sun, Ph.D. Candidate
Blekinge Institute of Technology
Karlskrona 37179, Sweden
bin.sun@bth.se

Wei Cheng, Ph.D. Prof.
Corresponding Author
Kunming University of Science and Technology
Kunming 650093, China
Blekinge Institute of Technology
Karlskrona 37179, Sweden
wei.cheng@bth.se

Guohua Bai, Ph.D. Prof.
Blekinge Institute of Technology
Karlskrona 37179, Sweden
guohua.bai@bth.se

Prashant Goswami, Ph.D. Assist. Prof.
Blekinge Institute of Technology
Karlskrona 37179, Sweden
prashant.goswami@bth.se