IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

Pattern Mining of Road Traffic in Developing Countries using Spatio-Temporal Data

Rajvi Kapadia[1], Varun Kasbekar[2], Vinaya Sawant[3]

Student, Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India [1, 2]

Assistant Professor, Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India[3]

Abstract: A lot of research has been done on city traffic routing mechanisms and congestion analysis using traffic sensors and CCTV cameras in spatio-temporal mining. Developing countries like India face periods of intensive rain for three-four months a year which has a drastic impact on road traffic. In this paper we have worked on creating a model for mining periodic patterns in traffic, which incorporates the fluctuations that occur due to monsoon in three-four months of the year, thereby depicting a holistic picture of traffic analysis in major cities of developing countries. The crucial aspect in Spatio-Temporal Data Mining is investigating temporal and spatial relations simultaneously as individual dimensions. For this reason it is important to choose a clustering algorithm which gives the most optimum performance for multi-dimensional datasets.

Keywords: Periodic patterns, Traffic Analysis, Spatio-temporal data, DB-Scan, Probability Distribution Matrices.

I. INTRODUCTION

The main issue in tracking traffic patterns in developing countries is lack of infrastructure like road traffic sensors and CCTV cameras in traffic junctions. Although there are cameras on national highways, it is difficult to assess the road traffic situation in the rest of the city. Data can be gathered using Google Maps API to give us a relative value of speeds in traffic junctions.
The ongoing research in this field fails to take into account the tropical monsoon climate of most of the developing countries like South East Asia, West and Central Africa and South west India. This climate deters us from having reliable predictive patterns throughout the year , having a huge impact on the traffic in large cities. In countries like India, that face the tropical monsoons from mid-June to mid –September there is large drop in vehicle speed in traffic junctions creating a havoc and major loss in productivity. This paper aims to incorporate the effect of the heavy monsoons on the road traffic and present a more accurate analysis of traffic patterns in highly populated developing cities like Mumbai.

Traffic pattern mining can be used to determine the best route to reach a particular destination and avoid major congestion areas. This route is determined taking into account the periodic speeds, time and weather conditions on the surrounding area. This can help vehicle users save time, increase productivity, and reduce stress levels.

II. LITERATURE SURVEY

The existing system addresses the issues of traffic management with feasible facilities available in urban areas like the large amount of traffic on urban roads, replacement of priority junctions with traffic signals etc. The most challenging aspect of traffic analysis is the prediction of traffic for detection of potential traffic jam spots. Using image processing and machine learning techniques they took aerial images and decision trees as input. Image processing techniques were used to map the geospatial data available in aerial images. A decision tree predicts the amount of traffic in the location with respect to time and date. The dataset used corresponded to transportation networks which contained the traffic state in various locations. The proposed system consisted of extracting geospatial data by utilising aerial images. Firstly, it converted color images into grayscale images. Then the system used global binarization to convert grayscale images into binary. Binary morphological operators were applied to the binary image , thereby removing obstacles and noise. In order to extract roads from the images they utilised “regionprops” function in MATLAB’s Image Processing Toolbox

The paper by Tjindal et all suggests discretisation of speed ranges into a speed time series which is then converted to a Boolean time series to detect periodicity in speed levels. They then find periods in the speeds using Discrete Fourier Transform to transform these speeds into complex numbers. DFT is a better bet than DTFT since it is completely discrete in terms of frequency and time. They then constructed a periodogram, which is a plot of power spectral density of the following complex numbers. Possible periods are detected from the periodogram by setting a threshold and highlighting all power densities above the same threshold. The next step is to use circular autocorrelation to identify true periods which lie on the hills of autocorrelation function as opposed to those who don’t. Periodic behaviour of stations is detected by creating a categorical distribution matrix. A categorical distribution matrix is created for every station using

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

maximum likelihood estimation. For the clustering of stations they used a metric which is derived from KL divergence as suggested by DominikEndres et al, called the Jersen-Shannon divergence (JS divergence) to obtain similarity measures. The stations with JS divergence less than the threshold were put together in a cluster. The threshold was set locally as compared to its neighbours because this helped discover clusters from neighborhoods depending on the average divergence between the neighborhood stations.

III. ISSUES IN SPATIO TEMPORAL MINING

Issues in spatio-temporal data mining are as follows:-

- Design analysis and validity of accurate spatiotemporal models and data structures is the most crucial issue for spatiotemporal data mining and data analysis.
- Carrying distance measures and topological information which require temporal and geometric computation are the salient features of spatiotemporal datasets.
- Visual representation of spatiotemporal patterns and schema, extent of scalability in data mining techniques, data structures to express robust methodology and index spatiotemporal sets of data are also alarming issues.
- Distance, topology, direction, climate are spatiotemporal parameters which provide specific information and are mandatorily needed to be considered in spatiotemporal data mining.
- Spatial and temporal inter-relationships are implicitly analysed and stated. They are not separately fed into or encoded in a database. These relationships must be evaluated and extracted from data.
- Scaling effect is a conflicting issue in spatiotemporal data mining. Temporal granularity or Spatial resolution with respect to time and space can have a direct impact on the efficiency and effectiveness of spatiotemporal relationships discovered in datasets.
- The standout feature of spatiotemporal datasets is that it requires specific modification or fabrication of data mining methods and procedures so that they can explore the rich spatial and temporal connections and patterns included in the datasets.
- The attributes of adjoining and neighbouring parameters can have some effect on the datasets which needs meticulous consideration. For example, spatiotemporal events like thunderstorm or cloudburst will have adverse effects on traffic patterns.
- Transitive and symmetric property in spatiotemporal data provides domain independent knowledge that should be taken care of while generating patterns. Implication of rules and standards and how to merge them with spatiotemporal data is a challenging issue.
- Features and functionalities should be explored to preprocess the data before mining and even at the time of computation when the data is actually needed.
- Development of effective visual interfacing techniques for viewing and accordingly modifying the geometrical and temporal attributes and parameters of spatiotemporal data is another important challenge.
- Patterns will be affected due to changes in the traits of adjacent patterns and should be reviewed. For example, spatiotemporal event like construction of buildings in forest regions will influence the movement pattern of the wild animals.

IV. PROPOSED SYSTEM

Since we cannot ignore the importance of having monsoon climate as an individual attribute while considering geographical distance and periodic speeds as part of clustering features, we propose a three-dimensional clustering method to incorporate the same. Here we shall consider geographical distance, periodic speed and rainfall as individual dimensions and implement DbScan clustering using Jersen-Shannon divergence (JSdivergence) similarity measures as a precomputed metric and visualization of the clusters using MDS(Multi-dimensional scaling algo) in 3-dimensional space.

V. CLUSTERING ALGORITHMS

A.DBSCAN

While clustering points are classified into core points, density reachable points and outliners.

Core points: They are said to be the points which have at least n(minpts) points within a maximum distance of Eps from them. These points are directly reachable to P.

Density reachable points: P is a point density reachable from Q when there are a set of points P1,P2…Pn where P1=P and Pn=Q, where each Pi+1 is directly reachable from Pi.

Density connected: Two points P and Q are said to be density connected if there is a point O, such that P and Q are density reachable from O.

Outliners: All points not reachable by any other points are outliners.

Cluster: “If point P is part of a cluster A and point Q is density reachable to Q with respect to toEps distance and minimum number of points, then Q is part of the cluster as well.”[3] Therefore the algorithm starts with a random point P and tries to find other points in a cluster with respect to MinPts and Eps distance. If P is a core point this algorithm will identify a cluster. If P is a border point it will not have density reachable points connected to it and the algorithm will move to the next point.

The DBScan algorithm has the following advantages:

- It can create clusters in a large spatial data set by identifying the local density of elements using a two input parameters- Eps distance and Minpts. Since the user can decide which parameter value is most suitable for Eps distance, very little domain knowledge is required.
- DbScan does not require one to determine number of clusters in advance like K-means.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 12, December 2016

- It can distinguish outliners from information and noise to a certain extent.
- DbScan scales almost linearly with the size of the database. Hence it is best for large spatial datasets.
- It can find clusters of an arbitrary shape. The link effect between closely spaced clusters is reduced by the minpts parameter.
- It is insensitive to the ordering of input data.

B. Clarans

Clarans algorithm is an improved K-medoids algorithm with improvement in efficiency for databases with about a thousand objects. When the number of objects increase, the runtime increases, wherein the algorithm lags behind. When compared to DBScan, CLARANS cluster objects with lower accuracy and runtime than DBScan. Data is taken from [5]

C.Naïve K-means

Naïve K-means partitions data sets into k divisions such that each division belongs to the same central point. It iteratively reallocates data into its respective divisions depending on the Euclidian distance to each centre such that points given in a particular subset are closer to their centre as compared to any other. The algorithm converges when reallocated data is allocated to same centres in the next iteration which indicates that this condition can always be satisfied. As compared to DBscan it is inefficient because k-means requires a huge number of nearest-neighbour queries depending on the number of neighbours(n) and dimensions(d); thus making it impractical to implement in a large dataset. It is also affected by the initial ordering of data in data sets unlike DbScan.

VI. SIMILARITY MEASURES

Kullback-Leibler divergence is the most popular similarity measures for probability distributions Jersen-Shannon divergence [8] is based on KL-divergence and converts it into a metric, making it a more suitable similarity measure for clustering.

VII. AVAILABILITY

The platform on which we implement clustering algorithm for large datasets is very crucial for its performance. The available tools for classification and clustering of datasets are as follows:

1.ELKI provides a powerful implementation of DBSCAN which leverages different index structures for sub-quadratic runtime and supports multi-dimensional data types. It might be outperformed by low-level optimized implementations on smaller data sets.
2.Scikit-learn provides a Python implementation of DBSCAN. Its ball-tree implementation supports a good selection of metrics. DBSCAN has been adapted to not compute the entire pairwise distance matrix internally, thereby optimising performance. Haversine distance can be used to optimally identify clusters.
3.Weka gives a basic implementation of DBSCAN that will run in quadratic time as well as linear memory.

VIII. CONCLUSION

Various traffic forecasting models have been developed using spatio temporal pattern mining. Their accuracy has improved manifold, but the factors considered must be modified in different locations. Tropical countries like India that face periodic rainfall must consider different factors while analysing and predicting future road traffic patterns. Optimising and combining different models is the developmental trend for pattern mining of road traffic using spatio temporal data.

REFERENCES

[1] Xiaolei Li, Jiawei Han, Jae-Gil Lee, and Hector Gonzalez. "Traffic density-based discovery of hot routes in road networks."Advances in Spatial and Temporal Databases (2007)
[2] Hector Gonzales, Jiawei Han, Xiaolei Li, Margaret Myslinska, and John Paul Sondag. "Adaptive fastest path computation on a road network: A traffic mining approach." In Proceedings of the 33rd international conference on Very large data bases, pp. 794-805. VLDB Endowment, 2007.
[3] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm to discover clusters in large spatial databases. AAAI Press
[4] Zhenhui Li, Jingjing Wang, &Jiawei Han (2012, August). Mining event periodicity from incomplete observations.In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 444-452).ACM.
[5] Mumtaz, K; Duraiswamy, K. An Analysis on Density Based Clustering of Multi Dimensional Spatial Data.
[6] Lu-An Tang, Yu Zheng, Jing Yuan, Jiawei Han, Alice Leung, Chih-Chieh Hung, and Wen-ChihPeng. "On discovery of traveling companions from streaming trajectories."In Data Engineering (ICDE), 2012 IEEE 28th International Conference on, pp. 186-197. IEEE, 2012.
[7] M. Vlachos, P. S. Yu, and V. Castelli.On periodicity detection and structural periodic similarity.In SDM, 2005.
[8] Félix Iglesias, and Wolfgang Kastner. "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns." Energies 6, no. 2 (2013): 579-597
[9] VenkateswaraRao,K; Govardhan A and ChalapatiRao K,V. “Spatiotemporal Data Mining:Issues, Tasks And Applications”.
[10] Prasad, KalliSrinivasa and Ramakrishna, Seelam “An Efficient Traffic Forecasting System Based on Spatial Data and Decision Trees”.