A Spatial Regression analysis Model for temporal Data mining in estimation of traffic data over a busy area

[1]A.V.N.Krishna, [2]M.Pavan Roy.
[1]Professor, Indur Institute of Engg. & Tech. Siddipet, Medak Dist. Andhra Pradesh, India.
[2]Asst.Professor, , Indur Institute of Engg. & Tech. Siddipet, Medak Dist. Andhra Pradesh, India.
Cell : 9849520995. Email: hariavn@yahoo.com

Abstract

Many applications maintain temporal & spatial features in their databases. These features cannot be treated as any other attributes and need special attention. Temporal data mining has the capability to infer casual and temporal proximity relationships among different components of data.

In this work a model is going to be developed which helps in measuring traffic data distributed over a wide area. This model considers the assumption that the data fallow ordered sequence. The area is divided into a set of grid points. Each grid point is identified by a set of coefficients. The traffic data at a set of locations is measured. The coefficients at the identified locations are calculated from the measured traffic data value. These coefficients are used to generate the traffic data at the other grid points by spatial regression analysis method. The procedure is repeated till the values cease to change for unit time. The procedure is repeated for different intervals of time. Thus traffic data is obtained over the wide area for different times and at different locations.

Key words

Spatial regression analysis , traffic data, temporal and spatial data, example.

1. Introduction

Temporal data mining is an important extension of data mining and it can be defined as the non trivial extraction of implicit, potentially useful and previously unrecorded information with an implicit or explicit temporal content from large quantities of data. It has the capability to infer casual and temporal proximity relationships and this is something non-temporal data mining cannot do. It may be noted that data mining from temporal data is not temporal data mining, if the temporal component is either ignored or treated as a simple numerical attribute. Also note that temporal rules cannot be mined from a database which is free of temporal components by traditional data mining techniques.

Regression analysis models the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands) (usually named Y), and the predictors (also called independent variables, explanatory variables, control variables, or regressors,) usually named $X_1,...,X_p$). If there is more than one response variable, it is called as *multivariate regression*.

2. Description

Types of temporal data,

1. Static: Each data item is considered free from any temporal reference and the inferences that can be derived from this data are also free of any temporal aspects.
2. Sequence. In this category of data, though there may not be any explicit reference to time, there exists a sort of qualitative temporal relationship between data items. The market basket transaction is a good example of this category. The entry sequence of transactions automatically incorporates a sort of temporality. If a transaction appears in the data base before

another transaction, it implies that the former transaction occurred before the latter. While most collections are often limited to the sequence relationships before and after, this category also includes the richer relationships, such as during, meet, overlap etc. Thus there exists a sort of qualitative temporal relationship between data items.

3. Time stamped. Here we can not only say that a transaction occurred before another but also the exact temporal distance between the data elements. Also with the events being uniformly spaced on the time scale.
4. Fully Temporal: In this category, the validity of the data elements is time dependent. The inferences are necessarily temporal in such cases.

3. Mathematical modeling of the problem

The approach to time series analysis was the establishment of a mathematical model describing the observed system. Depending on the appropriation of the problem a linear or non-linear model will be developed. This model can be useful to analyze census data, land use data and satellite meteorological data. The example that is going to be considered in this model is traffic data. It is based on general assumption that the traffic density is maximum at the center of city and it is minimum or thin at the borders. And also the traffic increase depends on the present data. The model depends on external factors like the nature of the day (say working day), atmospheric conditions etc. which is represented by coefficients p0, p1, p2 and p3.

3.1 Description of the problem

The area considered in the problem is divided into set of grid points. In the present problem the path is divided to 21 grid points. Thus considering initial traffic of 30 people at each grid point uniformly distributed through out the area ie at t=0,T(I)=Y(I)=30 people, where I=1,2…21 grid points.

Dividing the path into M number of points. For each control volume,
Data leaving the control volume –
Data entering the control volume = Data gain per unit time in the control volume.

Multivariate： $Y = p0 + p1*X + p2*Z + p3*X*Z$

Y=Measured traffic density at different grid points .

X=Grid Points

Z=Time Span over different grid points.

4. Results

Distributed Traffic data at different grid points is considered for 2 hours. The length of the considered path is 5 Km. The path is divided to 20 control volumes with a distance of 250 meters between each grid point. The time interval considered for the given problem i.e. del T is 12 minutes. Initially the traffic data at each grid point is assumed to be 30.

Thus δX= 250 meters
δt = 12 minutes.
Total time (n) = 2 hours.
$T_1 = T_M$ =30 population.

Thus by suitably calculating coefficients, traffic data could be generated at different points and at different times from the generated model. The traffic data sample is taken from police control room at different grid points for different times. This sampled data is used to generate the coefficients by multivariate regression analysis model with generated data from the model. Thus known the coefficients, the traffic analysis could be made at different grid points at different times which helps in taking proper corrective measures in handling traffic through the area.

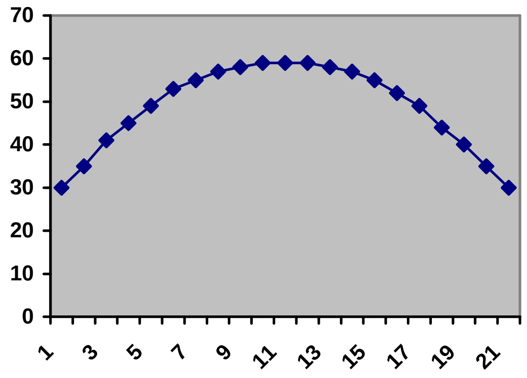
Traffic Data distribution over different grid points from records.

Grid Points	Traffic data	Grid Points	Traffic data
1	30	12	59
2	35	13	58
3	41	14	57
4	45	15	55
5	49	16	52
6	53	17	49
7	55	18	44
8	57	19	40
9	58	20	35
10	59	21	30
11	59		

p0=25.2 p1=3, p2=0.8 and p3=0.4

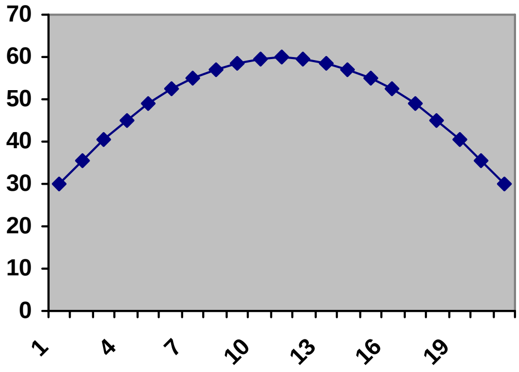
Traffic data distribution over different grid points from generated model.

Grid Points	Traffic data	Grid Points	Traffic data
1	30	12	59.5
2	35.5	13	58.6
3	40.6	14	57.1
4	45.1	15	55
5	49	16	52.3
6	52.3	17	49.2
7	55	18	45.1
8	57.11	19	40.6
9	58.6	20	35.5
10	59.5	21	30
11	59.8		



GENERATED DATA FROM THE MODEL

X axis –Grid Points.
Y axis –Traffic data.



SAMPLED DATA FROM RECORDS

5. Conclusion & Future work

In this paper, we propose a new method to identify incident in real time. It is based on spatial-temporal data view and applies tridiagonal matrix algorithm to consider the correlation of traffic data from neighboring stations and consecutive time slots. Our approach utilizes user feedback to support learning ability. Future efforts will be needed to refine the parameters in this approach. For incremental learning, re-calculating coefficients of each grid point costs extensive system resources. Incrementally update or approximate computing techniques can be applied to improve the computational efficiency of the proposed method. This approach can also be applied to other applications, which consider both temporal and spatial features, such as disease control and weather monitoring.

The model developed can be used for decision support systems of scientific computing. Thus some more applications of this model could be census data, land use data, satellite meteorological data and so on.

Future Work

Correlating behavior profiles across multiple links.
Integrate traffic profiling framework with a real-time monitoring system.

References

1. Dan.W.Patterson: *Introduction to Artificial Intelligence & Expert Systems*, Prentice-Hall of India Private limited –2001.
2. A.K.Pujari: *Data Mining Techniques*, Prentice-Hall of India Private limited-2002.
3. Ming-Syan Chen, Jong Soo Park, Philip S. Yu: Efficient Data Mining for Path Traversal Patterns. IEEE Trans. Knowl. Data Eng. 10(2): 209-221(1998)
4. Rakesh Agrawal, Christos Faloutsos, Arun N. Swami: Efficient Similarity Search In Sequence Databases. FODO 1993: 69-84 BibTeX
5. Frank Höppner. Learning Temporal Rules from State Sequences. In WLTSD, Seattle, USA, pages 25-31, 2001
6. Brockwell P.J. and Davis .R *Introduction to time series and forecasting*, Springer- Verlag, 1996.
7. Das G., Gunupolos D., Mannila *Finding similar time series*, Manuscrip. 1996
8. Faloutos C., Ranganathan M., Manolopoulos Y. *Fast subsequence matching in time series data bases*. In SIGMOD’94, 1994.
9. Ester M., Frommelt A., Kriegel H.P., and Sander J. *Algorithm for characterization and trend detection in* spatial data base, Fourth KDD conference,1998.
10. Allen J.F. *Maintaining knowledge about temporal intervals*, Commun.ACM, 26:11, 832-843, 1983.
11. Andrews,H.C.1972. *Introduction to mathematical techniques in pattern recognition.* New york, Wiley Interscience.
12. Brachman,R.J.1978 *A structural paradigm for representing knowledge*, Report No. 3605, Bolt Bernek and Newman ,Inc. Cambridge, Mass.
13. Duda,R.O., and P.E.Hart.1973, *Pattern Classification and Scene Analysis*, New York, Wiley.
14. Todorovski L, Dzeroski S, Srinivasan A, Whiteley J, Gavaghan D. Discovering the structure of partial differential equations from example behavior. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, 2000; 991-998.

- Abdi, H. "[1] *(2003). Partial least squares regression (PLS-regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795.]*".

- Abdi, H. "[2] *(2003). Partial regression coefficients. In M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795.]*".

- Abdi, H. "[3] *((2007). Coefficients of correlation, alienation and determination. In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage.*".

- Abdi, H. "[4] *((2007). Part and partial correlation. In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage.*".

- Abdi, H. "[5] *((2007). Multiple correlation coefficient. In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage.*".

Article received: 2007-05-21