Understanding City Dynamics by Manifold Learning Correlation Analysis

Wenzhu Zhang
Department of Electronic Engineering
Tsinghua University
zhwz@tsinghua.edu.cn

Lin Zhang
Department of Electronic Engineering
Tsinghua University
linzhang@tsinghua.edu.cn

ABSTRACT

Cities have long been considered as complex entities with nonlinear and dynamic properties. Pervasive urban sensing and crowd sourcing have become prevailing technologies that enhance the interplay between the cyber space and the physical world. In this paper, a spectral graph based manifold learning method is proposed to alleviate the impact of noisy, sparse and high-dimensional dataset. Correlation analysis of two physical processes is enhenced by semi-supervised machine learning. Preliminary evaluations on the correlation of traffic density and air quality reveal great potential of our method in future intelligent evironment study.

Categories and Subject Descriptors

I.2.6 [**Learning**]: [Knowledge acquisition]; I.4.8 [**Scene Analysis**]: [Sensor fusion]

1. INTRODUCTION

Sensors have become an essential part of urban lives. They extend human senses by revealing 'unseen' information layers. Trans-domain usage of sensor data and new emerging paradigms can change urban lives and policy-making processes by revealing hidden connections. Urban dynamics is the social behavior of human beings, which is known to relate strongly to environmental changes and resource consumption. City can be reasonably considered as an inherently human-driven self-organizing structure. Exploring the correlation between human activities, environmental change and resource consumption will be beneficial in many aspects including city plan, resource utility optimization, convenience improvement etc.

There has been a great shift by methodologies on urban dynamic study: from model-driven paradigm to data-driven paradigm[2]. The major challange now is how to perform efficient analysis with regard to oceans of data(if available) to obtain informative knowledge or prediction model. To be specific, three aspects should not be neglected.

1) *The unavoidable presence of noise or imprecision in training data adds uncertainty to the reconstruction process.*

2) *The sparsity of data obtained from crowd urban sensing cause incompleteness and heterogeneity of dataset both in space and time.*

3) *Quantitive analysis among different physical process in different measurement is difficult. Semantic abstraction are needed to gain meaningful information.*

Our goal is to develope proper methods which could alleviate the impact of noisy, sparse and high-dimensional dataset. We propose manifold learning based method to perform semantic abstraction, as well as spatial-temporal correlation to understand the implicit relationship between two phenomena.

2. SEMANTIC ABSTRACTION BASED ON MANIFOLD LEARNING

Suppose that in a city, there are N adjacent but not intersect areas $A_1, A_2, \ldots, A_N$. Each A_i, $i \in \{1, 2, \ldots, N\}$, contains m blocks $B_1, B_2, \ldots, B_m$. For every A_i, there are m measurement(sensory data) coming from m blocks respectively, denoted by $X_i(t)$. Define the learning result(output) as scalar $d_i^X(t)$. So we have we have m-dimesional input $X_i(t)$ and one scalar output $d_i^X(t)$ at time-step t in area A_i.

Suppose we have only partial knowledge about the input-output mapping. And there are two inter-related process $X_i(t)$ and $Y_i(t)$, $i \in \{1, 2, \ldots, N\}$. If we want to explore the implicit relationship between them, traditional methods adopt statistical methods such as canonical correlation analysis(CCA). However, it is hard to justify the meaning or significance of study results. Here we propose a new paradigm to perform correlation analysis. Firstly, for each dataset, we adopt manifold learning methods to reduce the dimension of data and obtain more 'abstract' information, which we could interperet as semantic level knowledge. Then spatial-temporal analysis could be used to gain higher level knowledge or prediction model. This approach is illustrated by Fig. 1.

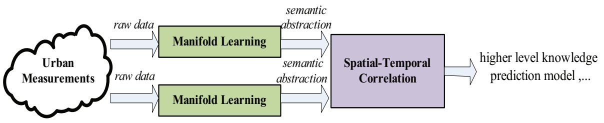


Figure 1: Manifold Learning Correlation Analysis

To overcome the problem of overfitting by a learning machine, regularization is usually used to restrict the solution of the hypersurface reconstruction problem by minimizing the augmented cost function. Generalized regularization theory extend classical theory by incorporating a second penalty function that reflects the intrinsic geometric structure of the input space. To be specific, the augmented cost function can

IPSN'12, April 16–20, 2012, Beijing, China.
ACM 978-1-4503-1227-1/12/04.

be expressed by

$$\Psi(F) = \Psi_s(F) + \lambda_A \Psi_c(F) + \lambda_I \Psi_I(F) \qquad (1)$$

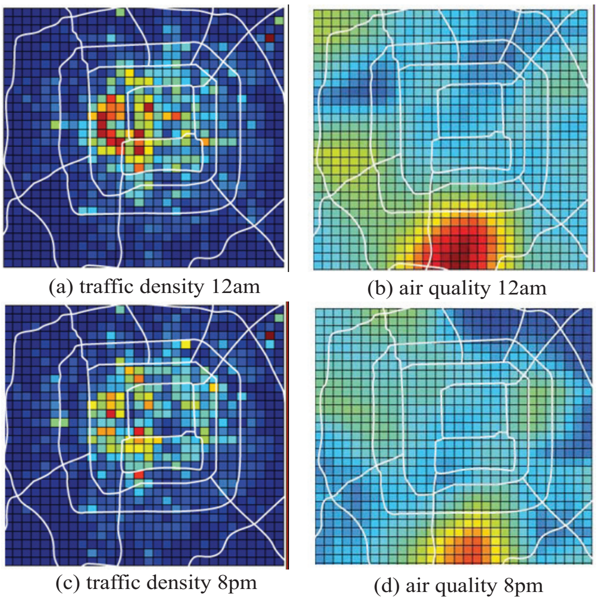
Here, $\Psi_s(F)$ denotes empirical cost function. While $\Psi_c(F)$ denotes regularizer, which is dependent on certain geometric properties of the approximating fuction. Thus, we propose a manifold based method to find proper $\Psi_I(F)$, which implies the intrinsic geometric structure of the input space. Here, we pursue the kernel approach based on manifold regularization. We use spectral graph theory to model a manifold.

Given this training sample, we proceed by constructing a weighted undirected graph graph consisting of N vertices, one for each input data point, and a set of edges connecting adjacent vertices. Let any two nodes i and j are connected, provide that the Euclidean distance between their respective data point $\mathbf{x}_i$ and $\mathbf{x}_j$ is small enough. Let w_{ij} denote the weight of an undirected edge connecting nodes i and j. Hereafter, we refer to the undirected graph, characterized by the weight matrix $\mathbf{W}$, as graph G. Let $\mathbf{T}$ denote an N-by-N diagonal matrix whose ii-th element is defined by $t_{ii} = \sum_{j=1}^{N} w_{ij}$, which is called the degree of node i. We define the *Laplacian* of graph G as $\mathbf{L} = \mathbf{T} - \mathbf{W}$.

Let $\Psi_I(F) = \mathbf{f}^T\mathbf{L}\mathbf{f}$ in equation (1).Define the vector valued function $\mathbf{f}$ in terms of the training sample X:$\mathbf{f} = [F(\mathbf{x}_1), F(\mathbf{x}_2), \ldots, F(\mathbf{x}_N)]^T$. According to generalized representer theorem[1], optimization of the cost function $\Psi(F)$ admits the form $F(\mathbf{x}) = \sum_{i=1}^{N} a_i k(\mathbf{x}, \mathbf{x}_i)$, where $k(\cdot,\cdot)$ is Gaussian kernel function.

3. CORRELATION OF TRAFFIC DENSITY AND AIR QUALITY

There are two datasets that we used for analysis, within the range of 5th ring road of Beijing city (E116.209-E116.544, N39.76-N40.02). For traffic density, we use Beijing taxi dataset with involves totally more than 20,000 taxi trajectories in one month. For the air quality, we used the dataset from our prototype system[3].


(a) traffic density 12am

(b) air quality 12am

(c) traffic density 8pm

(d) air quality 8pm

Figure 2: Traffic Density and Carbon Distribution

Fig.2 shows the density of vehicles and carbon-monoxide levels, where each small cell denotes 1km×1km area. We can see from Fig.2 that in downtown (inside the 3rd ring road), the traffic density is usually higher than that of other places, with the west region's higher than the east region's in downtown. We can see an obvious 'hot zone' of air quality, which indicates severe air pollution in that region. We find that there are several chemical plants in the south of Beijing city, which are reasonably responsible for the local air pollution.

Fig.3 shows the learning results at selected area (Dong Tie Ying Bridge, a 9 km^2 region with center E116.43, N39.856). We use a real value as uniformed index to represent the learning outputs for this specific area. The blue real line denotes traffic density, while red dotted line denotes air quality. It is inferred that the air quality is probably influenced by population density. For the selected area, we can predict with confidence the air pollution peak will occur approximately three hours later after the rush hour.

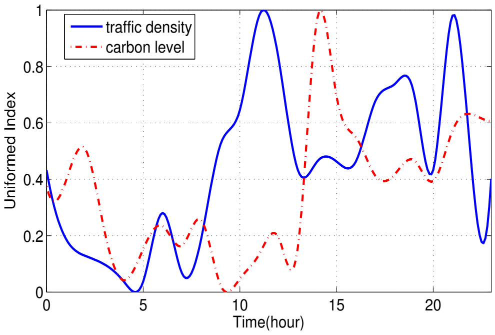


Figure 3: A Semantic Abstract of Traffic Density and Air Pollution in 24 hours

4. CONCLUSION

In this paper, we report our work progress on urban dynamics study. The major contribution of this paper is to use manifold learning on city phenomena correlation analysis. It reveals the intrinsic structure of dataset by spectral graph theory to achieve dimension reduction. In futhur study, spatial-temporal correlation methods can be developed to obtain non-trivial results. Interesting applications will be emerging towards better understanding of the cities.

5. REFERENCES

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, December 2006.

[2] I. Benenson. Modeling population dynamics in the city: from a regional to a multi-agent approach. *Discrete Dynamics in Nature and Society*, 3:149–170, 1999.

[3] W. Zhang, L. Zhang, Y. Ding, T. Miyaki, D. Gordon, and M. Beigl. Mobile sensing in metropolitan area: Case study in beijing. In *Mobile Sensing Workshop in 13th International Conference on Ubiquitous Computing (UbiComp'11)*, 2011.