

Finding Main Streets: Applying Machine Learning to Urban Design Planning

Jean Oh

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
jeanoh@cs.cmu.edu

Abstract. In the urban design realm careful consideration of connecting architectural form and socioeconomic function is a compelling issue. City planners and architects spend a significant amount of their time on collecting and integrating data from various information sources. Rapid growth of Geographic Information Systems (GIS) made it possible to map and display laboriously collected data, but these tools are limited by lack of sophisticated data analysis and inference capability. In this project we explored possibilities of how A.I. techniques can boost the performance of urban design planning by providing large scale data analysis and inference capability. Not to mention general benefits of automated process such as speed and labor cost, statistical analysis can also provide theoretical justification for designers which is not typically available in the case of manual efforts. As a proof of concept experiment we implemented an application of active learning that identifies a certain type of urban setting, Main Streets, based on their complicated spatial and semantic relationships over building geometry. The preliminary results show that active learning algorithm can effectively learn a classifier with relatively small number of training examples.

1 Introduction

City planners and architects spend a significant amount of their time on collecting, integrating, and analyzing information about urban components. Typically, data are scattered in various heterogeneous sources, e.g., building's spatial data in real estate web sites, tax data in city bureau reports documents, etc. Collected data are integrated based on the data point's geographic location as mapping keys, where Geographic Information Systems (GIS) are popularly used to suit this purpose. GIS provide an excellent way to map data from various sources using multi-layered infra-structure, supporting some limited capability of data analysis such as calculating mean and max. However, major part of semantic

analysis and construction of inference layers still remain to be manual work for human experts.

In this project we explored ideas of how A.I. techniques can assist designers in building complex urban models. Computer aided design is advantageous in terms of time and money, and it also produces theoretically justifiable models through statistical analysis. In fact, it is often a typical trend that inference made by human experts is not well explained in a clear way except trusting their artistic intuition. Ironically, because of the same reason it has been difficult to convince designers whether using A.I. techniques can truly be beneficial in this artistic and sensitive domain.

Our aim is not to replace human experts by computer designers but to “assist” the human designers by processing larger set of data at faster speed, presenting helpful insights that are difficult to oversee without the help of computational power. Among many potential applications we took a *typology* as our starting efforts which is an architectural terminology corresponding to a *classification* problem. As a proof of concept experiment we have implemented an application of active learning technique for classifying a certain type of urban setting, Main Streets [5][6], based on their complicated spatial and semantic relationships. The preliminary results show that typology prediction is indeed a machine learnable problem. Furthermore, we show that learning performance is significantly improved by using active learning algorithm.

2 Background: Urban Morphology

This section describes a little bit of background on the urban design domain and the obstacles that urban designers often face with, particularly issues of *urban morphology*. Urban morphology is a traditional and critical speculation of urban studies. It represents a city and its architectural components such as buildings and streets mainly from two separate but related perspectives: *built form* and *functional performance*.

The integrated perspective of form and function in urban studies is not an innovative notion. In fact, it has been the core subject of urban matters for a long time, but relatively few practical approaches have been developed yet. Rather, previous approaches mostly focused on one dominant aspect of either form or function from a particular view point, e.g. Architecture or Socioeconomics let alone. Furthermore, the range and definition of form and function varies according to diverse disciplines. For instance, while architects regard form as three dimensional shape of space and building components in the intimate level, economists as rather two dimensional shape of cartographic plane in the regional or national scale. While architects consider function as activities in individual buildings and

the in-betweens, policy makers as performance of parcel or zone in the whole system of the city.

In the urban design realm, sophisticated consideration of connecting architectural form and socioeconomic function is a compelling issue beyond allocation of certain types of buildings to land use zone. Traditionally, urban designers have used maps in their own ways in order to compile heterogeneous information and exhibit their interpretation. Currently, Geographic Information Systems (GIS) have been rapidly growing data infrastructure throughout majority of city subjects: planning, transportation, tax assessment, facility management, security, rescue, etc.

Over the past ten years the growth of the internet and computational affordability have brought immense convenience to many fields, and urban design community was not an exception. A lot more information of various levels became publicly available to designers. Despite overwhelmingly large amount of data, however, they hardly fit on the needs of urban design because 1) integrating data from heterogeneous information sources is difficult and time consuming, 2) GIS use relational database for their underlying structure and are inefficient to represent detailed architectural form, 3) there is no standardized GIS data format upon which everyone has agreed, and 4) Data analysis and inference still fully remain as human experts' responsibility.

Given this brief background of urban design problems the issues listed above certainly introduce many interesting A.I. research problems, such as data integration, data mining, knowledge representation, and machine learning. In the following sections we describe our initial efforts of applying machine learning to architectural typology problem.

3 Experiment: Finding Main Streets

3.1 Main Streets

The concept of *Main Streets* are introduced from the city revitalization projects dated back in 1970s, which was an attempt to identify commercial districts that have potentials for revitalization. The idea was to combine historic preservation with economic development to restore prosperity and vitality to downtowns and neighborhood business districts. The criteria of choosing a right commercial district varies from city to city, thus it is hard to find a generalized set of rules to distinguish Main Streets from rest of districts. Since one cannot apply one standard that works on a city to another city a whole new analysis of the new city from scratch is unavoidable.

Obviously, this is an expensive and time consuming process, and it is true not only in the Main Streets case but also true in the field of architectural typology in general. For instance, the ARTISTS (Arterial Streets Towards Sustainability) project in Europe was developed to identify types of streets in order to provide better insights to urban planners and economists. This 2.2 billion euros budget project involved 17 European countries and took three years to classify five categories of streets [8]. Their main effort was made to statistically analyze the characteristics of street functions and draw a two-dimensional classification table. This project was completed by human experts. Their experimental results include how they classified 48 streets into 5 categories based on their two-dimensional classification table.

We attempt to carry out similar classification tasks but in an automated way. We also propose our framework to be general so that it can easily be applied to new cities. In this project we implemented a general learning system that can identify Main Streets from data extracted from GIS.

3.2 Data preparation

Supervised machine learning techniques have been successfully applied in various domains such as text categorization [10]. Most of machine learning algorithms expect data to be a well defined set of tuples, but in reality this is rarely the case. For example, if data is stored in relational database with multiple tables the data must be preprocessed into a giant single table. Building inference network from relational database is an interesting area of research [2] and we also anticipate our future work more towards this direction.

In this experiment we preprocessed our data into a suitable form for general classification algorithms. We used the city of Boston Main Streets data (Figure 1) which is an ideal test bed for evaluation because 19 district were readily identified as Main Streets by field experts. We exported relational database tables from the GIS data that is available from the city of Boston and turn them into a set of tuples. Initially we started with two database tables: buildings layer and parcels layer. Note that the raw data is in building or parcel level whereas our target concept, Main Streets, is defined in district which is usually composed of several hundreds of buildings.

First step is to cluster buildings into a set of candidate districts and we implemented a simple data preprocessor for this purpose. Since Main Streets are organized as commercial district we first clustered commercial buildings within certain proximity boundary. Small clusters that have less than 10 commercial buildings were filtered out in this step. Second step is to widen the proximity boundary in order to include nearby buildings and parcels as part of selected districts. The number of buildings in resulting district candidates varied from



Fig. 1. Main Streets in Boston, Massachusetts

tens through hundreds. The proximity boundary limit was chosen empirically to generate reasonable size clusters. In order to set more correct cluster boundaries we will need to incorporate more separator data, e.g., geographic obstacles such as mountains or rivers, or man-made obstacles such as bridges and highways. We leave this part for a future work. Once we have the clusters we used aggregated data, such as average size of buildings, as the set of features. Although the raw data consists of more than 90,649 buildings and 99,897 parcels (total around 180,000 data points) preprocessed data produced quite small data set, only about 80 district candidates.

3.3 Learning algorithms

After the data is preprocessed into a set of tuples we learned binary classifiers on them. We tried k-Nearest Neighbor (kNN) classifier, Naive Bayes classifier, Decision Trees, and SVM. Among all SVM performed best, but Decision Trees were preferred by designers due to its comprehensible nature. Labeling is an expensive process in this domain because labeling one district requires complex and careful inspection of data and also involves field study. This cost-bounded domain constraint leads us to favor learning algorithms that works with relatively small number of training examples.

One such idea is active learning in which learning system actively chooses the next training example to be labeled. We took Tong and Koller’s [9] approach over SVM. The basic idea is to suggest data points that are near the separation boundary, which is quite intuitive and also proven to be very effective in other practical domains such as text classification.

Semi-supervised learning utilizes distribution of large inexpensive unlabeled data to guide supervised learning. Particularly, co-training [1] learns two classifiers using disjoint sets of features, i.e., two different views over the same data, and admit the predictions upon which both classifiers agree.

More recent approach includes incorporating clustering into active learning [7]. Using prior data distribution their system first clusters data and suggests cluster representatives to active learner. Their algorithm selects not only the data points close to classification boundary but also representatives of unlabeled data. We adopted their idea to find the initial samples to be labeled. Since the size of our experimental data set was small performance was more sensitive to training examples at early learning steps. Preclustering was not very helpful in the later learning steps because the size of unlabeled data was not large enough. (After data preprocessing we had only about 80 district candidates)

4 Results and Conclusion

We explored possibilities of interesting A.I. research in urban planning domain. Since the urban planning community is conservative towards computational assistance our primary goal was to implement a simple example prototype to showcase potentials of applying machine learning and A.I. techniques. In our preliminary experiment of finding Main Streets, the results has two major contributions. First, we showed that architectural typology problem can be modeled as a classification problem except the fact that classification is more generalized paradigm. Second, using active learning our system can cleverly choose better samples to label, outperforming random selection model significantly as shown in Figure 3.

Figure 2 shows performance of using active learning in three different measures: precision, recall, and harmonic mean. In this domain precision is a more important measure than recall since Main Streets are eventually investment targets. For example, investing to a wrong district is much costly than missing one district. That being said, our system’s precision level is almost perfect after seeing 14 labeled examples. Harmonic mean also reaches over .9 after about 30 samples. When compared with random selection model, which chooses the next training example at random, the advantage of using active learning is obvious (Figure 3). The random selection model’s performance remains below .6 until it sees about

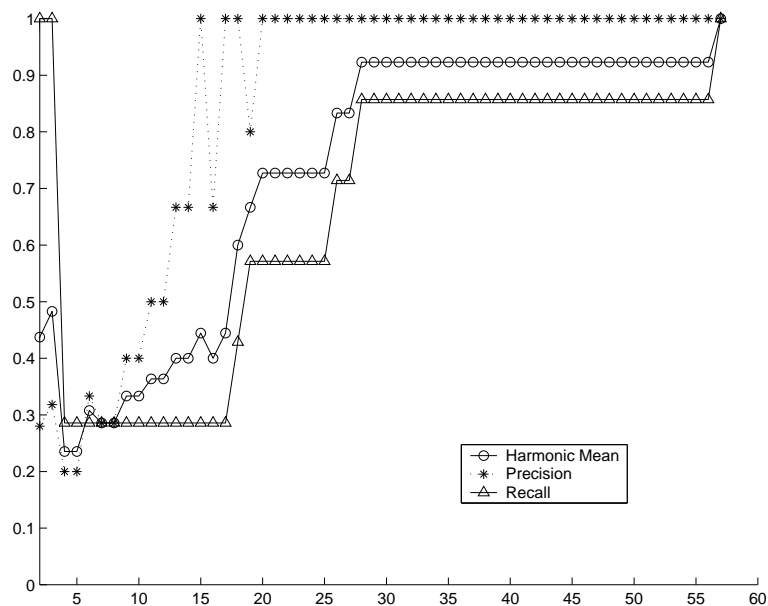


Fig. 2. Active Learning Result

45 training examples whereas the active learning model learns at fast speed from seeing 14-20 examples, achieving over .7 in harmonic mean.

In order to classify Main Streets designers normally use a lot more features than the ones used in our experiment. In fact, what we used was a very limited set of feature data. We expect learning speed to be accelerated when we use more feature data. Another noise factor was the district boundary. We roughly cut the buildings into a set of district blocks solely based on Euclidean distance. Despite the limitation of incomplete data and rough data preprocessing the result was reasonably good and convincing.

5 Future Work

Our on-going effort on the Main Streets project is to design a graphical user interface that designers can use for more comprehensive experiments. We also want to apply our framework to other cities whose structure are different from Boston.

In the bigger picture of urban planning domain, there are many more interesting A.I. research topics. The ultimate future goal is to construct architectural

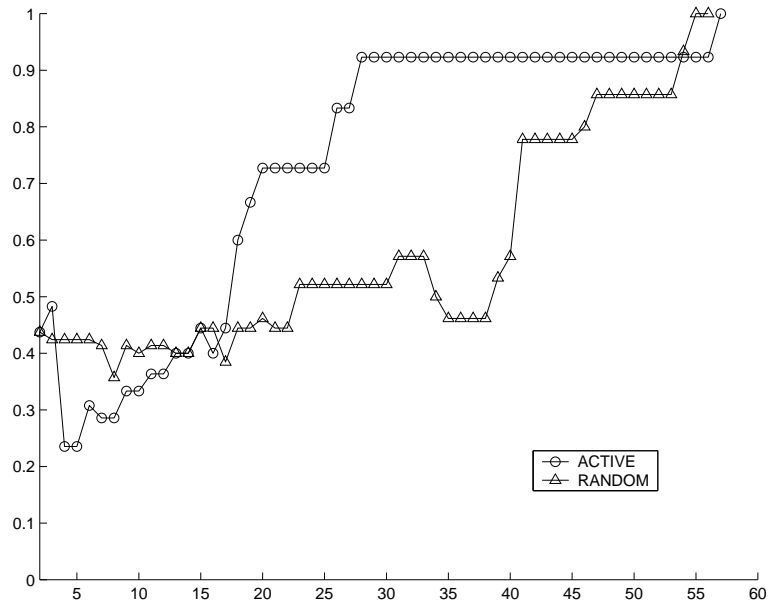


Fig. 3. Harmonic Mean: Random Selection vs. Active Learning

inference network on top of heterogeneous information sources which can assist urban planners by providing convenient data analysis and inference capability.

Data integration from heterogeneous information sources [4] is still a difficult problem. Particularly handling object identification problem (objects names are not unified in all sources) and data inconsistency are good examples of challenges in data integration area. Defining appropriate knowledge representation for architectural typology is also interesting. Previous work showed that XML-based languages are suitable for storing architectural metadata [3]. As we showed in our preliminary experiment there are a lot of opportunities in typology to use various machine learning techniques. We also plan to design a general intelligent user interface that seamlessly connects various underlying learning components into our urban planner assistant system.

6 Acknowledgement

This is a joint work with Jie-Eun Hwang at Graduate School of Design, Harvard University.

References

1. Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, pages 92–100, 1998.
2. Lise Getoor. *Learning Statistical Models from Relational Data*. PhD thesis, Stanford University, 2001.
3. J-E. Hwang and J-W. Choi. Spacecore: Metadata for retrieving spatial information in architecture. In *Proceedings of Association for Computer Aided Design in Architecture*, pages 24–27, 2002.
4. Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank. Mixed-initiative, multi-source information assistants. In *Proceedings of Tenth International World Wide Web Conference*, pages 697–707, 2001.
5. R. W. Longstreth. *The buildings of Main Street: a guide to American commercial architecture*. AltaMira Press, Walnut Creek, 2000.
6. L. Lopilato. *Main street: some lessons in revitalization*. , University Press of America, Inc., New York, 2003.
7. Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of International Conference on Machine Learning*, 2004.
8. Ase Svensson. Arterial Streets For People. Technical report, Lund University, Department of Technology and Society, Sweden, 2004.
9. Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In Pat Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 999–1006, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.
10. Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.