Predicting Land Use of Italian Cities using Structural Semantic Models

Gianni Barlacchi[1,2], **Bruno Lepri**[3], **Alessandro Moschitti**[1,4]
[1]Department of Information Engineering and Computer Science, University of Trento
[2] TIM Semantics and Knowledge Innovation Lab, Trento
[3] Fondazione Bruno Kessler, Trento
[4]Qatar Computing Research Institute, HBKU
{gianni.barlacchi,amoschitti}@gmail.com
lepri@fbk.eu

Abstract

English. We propose a hierarchical semantic representation of urban areas extracted from a social network to classify the most predominant land use, which is a very common task in urban computing. We encode geo-social data from Location-Based Social Networks with standard feature vectors and a conceptual tree structure that we call Geo-Tree. We use the latter in kernel machines, which can thus perform accurate classification, exploiting hierarchical substructure of concepts as features. Our comparative study on three datasets extracted from Milan, Rome and Naples shows that Tree Kernels applied to Geo-Trees are very effective improving the state of the art.

Italiano. *In questo lavoro, proponiamo un nuovo modello semantico per la rappresentazione di aree urbane utilizzando dati da social media. In particolare, modelliamo tale informazione con una struttura ad albero che abbiamo chiamato Geo-Tree. Questa viene utilizzata, in combinazione con un vettore di feature classico, nelle kernel machine per fare classificazione della destinazione di uso delle aree urbane. Abbiamo valutato il nostro approccio su tre grandi metropoli italiane quali Milano, Roma e Napoli. I risultati mostrano come i Geo-Tree, applicati ai Tree Kernel, riescono a raggiungere risultati di molto superiori ad altri modelli attualmente stato dell'arte.*

1 Introduction

The growing availability of data from cities (Barlacchi et al., 2015a) (e.g., traffic flow, human mobility and geographical data) opens new opportunities for predicting and thus optimizing human activities. For example, the automatic analysis of land use enables the possibility of better administrating a city in terms of resources and provided services. However, such analysis requires specific information, which is often not available for privacy concerns. In this paper we follow the approach proposed in (Barlacchi et al., 2017) and we use public textual descriptions of urban areas to design a novel machine learning representation. We represent urban areas as: (i) a bag-of-concepts (BOC), e.g., the terms *Arts and Entertainment*, *College and University*, *Event*, *Food* extracted from the Foursquare description of the area; and (ii) the same concepts above organized in a tree, which reflects the hierarchical organization of Foursquare activities. We combine BOC vectors with Tree Kernels (TKs) (Collins and Duffy, 2002; Moschitti, 2006) applied to concept trees (Geo-Tree) and use them in Support Vector Machines (SVMs). The Geo-Tree allows the model to learn complex structural and semantic patterns from the hierarchical conceptualization of an area. We show that TKs not only can capture semantic information from natural language text, e.g., as shown for semantic role labeling (Moschitti et al., 2008) and question answering (Severyn and Moschitti, 2013; Barlacchi et al., 2015b), but they can also learn from the hierarchy above to perform semantic inference, such as deciding which the major activity of a land is.

We carried out a study on land use prediction of three Italian cities: Milan, Rome and Naples as follows: (i) we divided each city in squares of 200x200 meters; (ii) then, we classify the most predominant land use class (e.g., *High Density Urban Fabric* or *Open Space and Outdoor*), assigned by the city administration. The results show that GeoTKs achieve an impressive improvement over state-of-the-art classification approaches based on BOC., i.e., 21.2%, 13.6% and 54.3% of relative improvement in Macro-F1 over Milan, Rome and

Naples datasets, respectively.

2 Related Work

Previous work has modeled land use classification by means of different sources of information. For example, Yuan et al. (2012) built a framework that, using human mobility patterns derived from taxi-cab trajectories and Point Of Interests (POIs), classifies the functionality of an area for the city of Beijing. Assem et al. (2016) proposed a spatio-temporal approach based on three different clustering algorithms to model the change of functionality of a city's region over time. They extracted features from Foursquare's POIs and check-in activities of Manhattan. Yao et al. (2017) built sequences of POI concepts reflecting their spatial distance. Then, they applied Word2Vec (Mikolov et al., 2013) to these sequences to derive vectors representing each area, which was used to train a land use classifier. In general, most previous work applies extensive feature engineering, which is typically costly as it requires to fully understand the target domain. Our approach alleviates this problem with automatic feature engineering applied to an abstract land representation.

3 Land Description Data

Geospatial city areas are described with the popular shape file format, where each shape is a collection of points geo-localized using their coordinates. The latter are provided with the well-known Coordinate Reference System (CRS) WGS84, adopted for the common latitude/longitude geolocation. We use (i) shape files provided by Urban Atlas[1], a website providing data for large urban areas (more than $100,000$ inhabitants) and (ii) POIs from Foursquare[2].

3.1 Land Use

Cities are divided in small areas associated with a main land use. In total, there are 17 different land use classes defined from the open dataset Urban Atlas [3]. We focused on those related to city centers, discarding those less interesting from a social viewpoint, i.e., associated with rural areas such as forests, agricultural, semi-natural and wetland areas and mineral extraction and dump sites. Thus, we selected the following categories: (i) *High Density Urban Fabric*, (ii) *Medium Density Urban Fabric*, (iii) *Low Density Urban Fabric*, (iv) *Industrial, commercial, public, military and private units*, (v) *Open Space & Recreation*, (vi) *Transportation*. We collapsed *Medium* and *Low Density Urban Fabric* into one single category, *ML-Density Urban Fabric* as they only have few samples. Land use distribution is very fine-grained, making its classification based on POI information very difficult. A trade-off between classification accuracy and the desired area granularity consists in segmenting the regions in squared cells. As each cell can contain more than one land use label, we consider the predominant label as its primary use.

3.2 Point-Of-Interest

A POI is usually characterized by a location (i.e., latitude and longitude), textual information (e.g., a description of the activity in that place) and a hierarchical categorization that provides different levels of detail about the activity of the place (e.g., *Food*, *Asian Restaurant*, *Chinese Restaurant*). We used POIs extracted from Foursquare, a geolocation-based social network supported with web search facilities for places and a recommendation system. In particular, we extracted 46,731, 43,389 and 7,219 POIs from Milan, Rome and Naples[4], respectively. We focused on the ten macro-categories of such POIs[5], each one specialized in maximum four levels of detail.

4 Structural Models

In most machine learning algorithms data examples are transformed in feature vectors, which in turn are used in dot products to carry out both learning and classification. Kernel Machines (KMs) allow for replacing the dot product with kernel functions, which directly compute it on the examples, i.e., they avoid the transformation of examples in vectors. The main advantage of KMs is a much lower computational complexity as it does not directly depend on the feature space size.

4.1 Point-of-interests Features

The most straightforward way to represent an area by means of Foursquare data is the use its POIs. Every venue is hierarchically categorized (e.g., *Professional and Other Places* $\rightarrow$ *Medical Center* $\rightarrow$ *Doctor's office*) and the categories are used to produce an aggregated representation of the area.

[1] https://www.eea.europa.eu/data-and-maps/data/urban-atlas

[2] https://foursquare.com/

[3] https://www.eea.europa.eu/data-and-maps/data/urban-atlas#tab-additional-information

[4] For some reasons Foursquare is less popular in Naples

[5] https://developer.foursquare.com/categorytree

We define a feature vector for a grid cell by counting the macro-level category (e.g., *Food*) in all the POIs that we found in that cell.

4.2 Geographical Tree Kernel

Foursquare has its own hierarchy of categories, which is used to characterize each location and activity (e.g., restaurants or shops) in the database. Thus, each Foursquare POI is associated with a hierarchical path, which semantically describes the type of location/activity (e.g., for *Chinese Restaurant*, we have the path *Food* $\rightarrow$ *Asian Restaurant* $\rightarrow$ *Chinese Restaurant*). The path is much more informative than just the target POI name, as it provides feature combinations following the structure and the node proximity information, e.g., *Food & Asian Restaurant* or *Asian Restaurant & Chinese Restaurant* are valid features whereas *Food & Chinese Restaurant* is not.

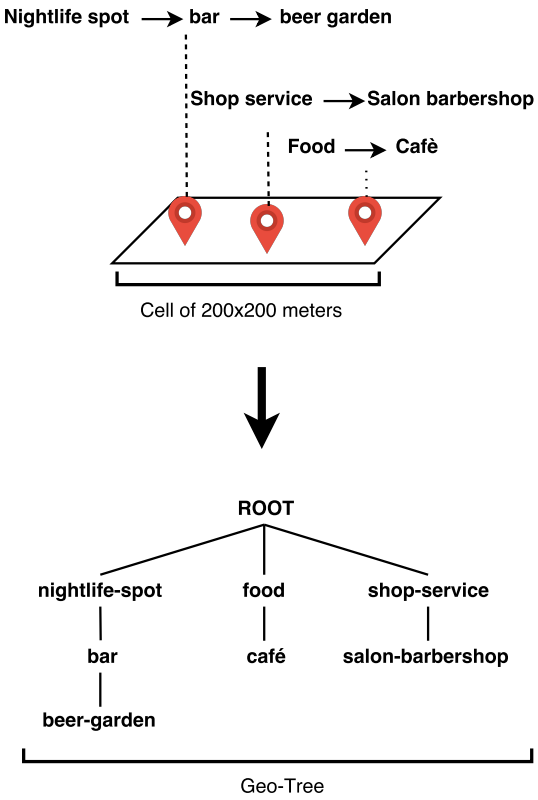

Figure 1: Example of Geo-Tree built from a collection POIs in a cell.

Geo-Tree: we propose a new tree structure, i.e., Geo-Tree, whose nodes and edges among them are subsets of the Foursquare hierarchy (FH). A Geo-Tree of a grid cell is constituted by a new root node connecting the subtrees of FH rooted in concepts present in the cell. In other words, we connect all the paths of FH starting from grid concepts. Figure 1 shows an example of the FH paths of a cell and the resulting Geo-Tree.

This way, the nodes of the first level, i.e., the root children, correspond to the most general FH categories, e.g., *Arts & Entertainment*, *Event*, *Food*, etc., the second level of our tree corresponds to the second level of the hierarchical tree of Foursquare, and so on. The terminal nodes are the finest-grained descriptions in terms of category about the area, e.g., *College Baseball Diamond* or *Southwestern French Restaurant*. For example, Fig. 2 illustrates the semantic structure of a grid cell obtained by combining all the categories' chains of each venue.

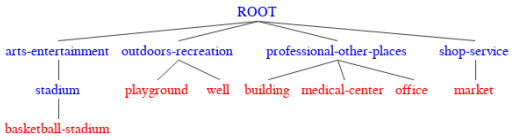

Figure 2: Example of Geo-Tree in Milan for an area labeled as Open Space & Recreation.

GeoTK: given a Geo-Tree, we can encode all its substructures in kernel machines using TKs. In particular, we used the Syntactic Tree Kernels (STK$_b$) with Bag-Of-Words and the Partial Tree Kernel (PTK) (Moschitti, 2006). Our TKs by construction do not consider the frequency[6] of the POIs present in a given grid cell.

BOC kernel: to complement GeoTK, we represent a cell also creating a BOC representation, namely we count the macro-level category (e.g., *Food*) in all the POIs that we found in any cell grid. This way, we generate feature vectors by counting the number of each activity under each macro-category. In order to take the popularity of the area into consideration, we included (i) the total sum of unique users that did at least one check-in in the cell, and (ii) the total sum of check-in done in the cell. Note that, given an area, the number of unique users provides an idea on how many people visited it, while the number of check-in can be used to represent its popularity.

Kernel combination: finally, given two geographical areas, x^a and x^b, we define a kernel combining Geo-Tree and BOC as: $K(x^a, x^b) = TK(\mathbf{t}^a, \mathbf{t}^b) + KV(\mathbf{v}^a, \mathbf{v}^b)$, where TK is any

[6] It is possible to add the frequency in the kernel computation but for our study we preferred to have a completely different representation from previous typical frequency-based approaches.

structural kernel function applied to tree representations, $\mathbf{t}^a$ and $\mathbf{t}^b$ of the geographical areas and KV is a kernel applied to the feature vectors, $\mathbf{v}^a$ and $\mathbf{v}^b$, extracted from x^a and x^b using any data source available (e.g., text, social media, mobile phone and census data).

5 Experiments and Results

We performed our experiments on the data from Milan, Rome and Naples. We used a grid of 200x200meters as it is indicated as the best size from other similar previous work on land use classification (Toole et al., 2012; Zhan et al., 2014; Barlacchi et al., 2017). We applied a preprocessing step in order to filter out cells for which land use classification cannot be performed. In particular, for Milan and Rome, we selected the central point of the shape and we included those cells that have their centroid in the radius of 15 and 8 kilometers, respectively. For Naples, we kept all the cells due to the smaller size of the city. Then, for all the three cities, we removed the cells that (i) cover areas without a specified land use (e.g., the cells in the sea) and (ii) do not have POIs (e.g., the countryside cells). After this step, we obtained a grid with 2,581, 5,657 and 1,314 cells for Milan, Rome and Naples, respectively. We created, separately for each city, the training and test set randomly sampling 80% vs. 20% of the cells. We labelled the dataset following the same category aggregation strategy proposed by Zhan et al. (2014), who assigned the predominant land use class to each grid cell.

To train our models, we applied SVM-Light-TK[7], which enables the use of structural kernels (Moschitti, 2006) in SVM-Light[8]. In particular, due to the nature of the task, we used a Python wrapper around SVM-Light-TK to perform multiclass classification[9]. We experimented with linear, polynomial and radial basis function kernels applied to standard feature vectors. We measured the performance of our classifier by averaging Precision, Recall and F1 over all land use categories.

5.1 Results for Land Use Classification

We trained multi-class classifiers using common learning algorithm such XGboost (Chen and Guestrin, 2016), and SVM using linear, polynomial and radial basis function kernels, named SVM-{Lin, Poly, Rbf}, respectively, and our structural semantic models, indicated with STK_b and PTK. We also combined kernels with a simple summation, e.g., PTK+Lin indicates an SVM using such kernel combination.

City	Model	Prec.	Rec.	F1
Milan	baseline	0.200	0.119	0.149
	XGBoost	0.294	0.317	0.297
	STK_b+Rbf	0.368	**0.364**	**0.360**
	PTK+Rbf	0.430	0.350	0.345
	STK_b	**0.448**	0.307	0.320
	PTK	0.364	0.302	0.309
Rome	baseline	0.200	0.089	0.124
	XGBoost	0.291	0.306	0.279
	STK_b+Lin	**0.359**	**0.314**	**0.317**
	STK	0.338	0.300	0.302
	PTK	0.340	0.300	0.299
	PTK+Lin	0.359	0.297	0.291
Naples	baseline	0.200	0.100	0.133
	XGBoost	0.236	0.272	0.219
	STK_b+Rbf	**0.361**	**0.331**	**0.338**
	STK_b+Lin	0.338	0.302	0.300
	STK_b	0.409	0.290	0.299
	PTK	0.318	0.298	0.297

Table 1: Classification results on Rome, Milan and Naples. Prec., Rec. and F1 are averaged over all categories.

Table 1 shows the average of F1, Precision and Recall over the different categories. The model *baseline* is obtained by always classifying an example with the label *High Density Urban Fabric*, which is the most frequent. Due to space constraint, we only reported six models, namely: the baseline, XGBoost and the top four kernel models.

We note that: (i) GeoTK always outperforms XGBoost and the baseline, demonstrating the superiority of our novel approach. This is an interesting finding as XGboost is the current state of the art for land use classification. (ii) STK_b combined with feature vector always produces the best results, improving the F1-score over XGBoost up to 6.3, 3.8 and 11.9 absolute points for Milan, Rome and Naples, respectively. (iii) Kernel combinations always provide the best results.

6 Conclusions

In this paper, we have introduced Geo-Trees, a novel semantic representation based on a hierarchical classification of POIs, to better exploit geo-social data to the classification of the primary land use of an urban area. This is an important task as it gives the urban planners and policy makers the possibility to better administrate and renew a city in terms of infrastructures, resources and services. More in detail, we have built our classifiers with combinations of a kernel over BOC and TKs applied to Geo-Trees, thus exploiting hierar-

[7] http://disi.unitn.it/moschitti/Tree-Kernel.htm

[8] http://svmlight.joachims.org/

[9] https://github.com/aseveryn/SVMTK-Multiclass-Classifier

chical substructure of concepts as features. Our comparative study on three large Italian cities, Milan, Rome and Naples shows that our models can relatively improve the state of the art up to 11.9 absolute points in F1-score.

Acknowledgments

This work has been partially supported by the EC project CogNet, 671625 (H2020-ICT-2014-2, Research and Innovation action).

References

Haytham Assem, Lei Xu, Teodora Sandra Buda, and Declan O'Sullivan. 2016. Spatio-temporal clustering approach for detecting functional regions in cities. In *ICTAI*, pages 370–377. IEEE.

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015a. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2:150055.

Gianni Barlacchi, Massimo Nicosia, and Alessandro Moschitti. 2015b. Sacry: Syntax-based automatic crossword puzzle resolution system. *ACL-IJCNLP 2015*, page 79.

G Barlacchi, A Rossi, B Lepri, and A Moschitti. 2017. Structural semantic models for automatic analysis of land use.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, New York, NY, USA. ACM.

Michael Collins and Nigel Duffy. 2002. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.

Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *ECML*, pages 318–329. Springer.

Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic feature engineering for answer selection and extraction. In *EMNLP*, volume 13, pages 458–467.

Jameson L Toole, Michael Ulm, Marta C González, and Dietmar Bauer. 2012. Inferring land use from mobile phone activity. In *SIGKDD International Workshop on Urban Computing*, pages 1–8. ACM.

Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848.

Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and pois. In *KDD*, pages 186–194. ACM.

Xianyuan Zhan, Satish V Ukkusuri, and Feng Zhu. 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics*, 14(3-4):647–667.